LEVEL ~~III~~ A072759

AD A072760

Proceedings of The Pacific Conference on Operations Research

April 23 – 28, 1979, Seoul, Korea

VOL. II

THE PACIFIC CONFERENCE 1979

PUBLISHED BY
Military Operations Research Society of Korea
Korean Operations Research Society

Printed by Young Jin Sa, 342, 3Ka, Ulchi-Ro, Chung-Ku, Seoul, Korea.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Proceedings of the Pacific Conference on Operations Research, April 23-28, 1979, Seoul, Korea VOL II		5. TYPE OF REPORT & PERIOD COVERED Final 23-28 April 1979
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dr. Rak To Song, Prof Soondal Park, and Dr. Ui Chong Choe (Editors)		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Military Operations Research Society of Korea Korean Operations Research Society		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Korea Institute for Defense Analyses C.P.O. Box 3089 Seoul, Republic of Korea		12. REPORT DATE April 1979 13. NUMBER OF PAGES 814
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

See also VOL I and Addendum to the Proceedings.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Operations Research	Pacific Area	Cost Effectiveness
Systems Analysis	Korea	Mathematical Modelling
Conference	Weapons Systems Evaluation	
Symposium	Cost and Operational Effectiveness	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

These Proceedings and the Addendum contain the texts of the keynote speeches, the state-of-the-art lectures, and the papers presented at the Pacific Conference on Operations Research, held April 23-28, 1979, in Seoul, Korea. Analysts and other professionals representing government agencies, academic institutions and industry from 18 countries participated, and a total of 71 presentations were published in the two volumes of the Proceedings and the Addendum.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(9) Final rept.,

LEVEL II

(2)

(6) PROCEEDINGS

OF

THE PACIFIC CONFERENCE

ON

OPERATIONS RESEARCH

VOL. II

held

APRIL 23-28, 1979,

SEOUL, KOREA. Volume II.

(11) Apr 79

(12) 730p.

Accession For	
NTIS GRA&I	X
DDC TAB	
Unannounced	
Justification	
By	
Distribution/	
Availability Codes	
Dist.	Avail and/or special
A	

(10) *EDITED BY*

DR. RAK TO SONG
PROF. SOONDAL PARK
DR. UI CHONG CHOE

1B

This Conference was

Sponsored by

The Military Operations Research Society of Korea
The Korean Operations Research Society

in Collaboration with

The International Federation of Operational Research Societies

ORGANIZATION OF THE CONFERENCE

Chairman:

Dr. Moon Taik Shim
President, Military Operations Research Society of Korea

Co-Chairman:

Mr. Eung Kyun Shin
President, Korean Operations Research Society

Chairman of the Organizing Committee:

Maj. Gen. Jang-Nai Sohn
Vice President, Military Operations Research Society of Korea

Vice Chairmen:

Dr. Sung Jin Kim, Korea Institute for Defense Analyses

Brig. Gen. Yon Sik Choi
Secretary General, Military Operations Research Society of Korea

Secretary General:

Dr. Rak To Song, Korea Institute for Defense Analyses

Planning Committee:

Dr. Man Suk Song,
Mr. Hee-Myon Kwon, and
Mr. Kil Ho Chung, Korea Institute for Defense Analyses
Prof. Hyung Jae Oh, City University of Seoul

Editorial Committee:

Dr. Soon Dal Park, Seoul National University
Dr. Ui Chong Choe, Republic of Korea Navy

Publicity Committee:

LTC In Soo Kang, Republic of Korea Navy

Protocol Committee:

Mr. Kyu Pok Lee, Agency for Defense Development

Budget Committee:

Mr. Yang Il Sin, Agency for Defense Development

PREFACE

The Pacific Conference on Operations Research, held from 23 to 28 April, 1979, was sponsored by the Military Operations Research Society of Korea and the Korean Operations Research Society in collaboration with the International Federation of Operational Research Societies. These Proceedings contain the texts of the state-of-the-art lectures and papers presented at the Conference. We are most grateful to the authors and to the session chairmen for their efforts in bringing about the success of the Conference, and hope that the excellent papers published in these proceedings further stimulate the international exchange of ideas, to the benefit of all participants.

We gratefully acknowledge the financial support provided by the Ministry of National Defense, Republic of Korea, and the cooperation of the International Federation of Operational Research Societies. We are also indebted to Admiral Carlisle Trost, Mr. John Gratwick, and Dr. Joseph Sperrazza for their stimulating keynote addresses, and to Prof. B. Hutchinson, Dr. D. Schrady, Mr. Charles Wolf, Prof. S.M. Lee, Prof. T. Nishida, Dr. D. Hirshfeld, and Dr. Kong-Kyun Ro for their excellent state-of-the-art lectures. Last but not least, special thanks are due to organizing committee members Dr. Rak To Song, Dr. Soondal Park, Dr. Ui Chong Choe, Dr. Man Suk Song, Mr. Hee Myon Kwon, Mr. Kil Ho Chung, Prof. Hyung Jae Oh, Mr. In Soo Kang, and Mr. Juri Toomepuu, whose selfless contributions to the success of the Conference are immeasurable.

Moon Taik Shim, Ph.D.
Chairman

April 1979
Seoul, Korea

CONTENTS

CONTENTS OF VOL. II

Session V : Business Management (I)

Session VI : Theory of Operations Research

Session VII : Military Operations Research (III)

Session VIII : Business Management (II)

Session IX : Public Administration

CONTENTS

CONTENTS OF VOL. I

The State-of-the-Art Lecture

Session I : Transportation Systems Analysis

VOL. II

SIMULATING STRUCTURED SCENARIOS FOR CORPORATE PLANNING

John Tydeman
Australian National University
A.C.T., Australia, and
Institute for the Future
Menlo Park, Cailfornia, U.S.A.

Robert Mitchell
Canberra College of Advanced Education
A.C.T., Australia

ABSTRACT. The operations research contribution to both public- and private-sector planning has become increasingly significant over the last decade. In particular, the need to formally include an uncertainty component in any planning model is widely accepted. An operational procedure, which gives the decision maker the flexibility to plan in a changing environment and to evaluate alternative strategies, currently lacks formal definition.

The aim of this paper is to outline a method of generating a structured uncertainty context for use in problem analysis in general and corporate planning in particular and to report its application to a large Australian public authority. The uncertainty context is defined to comprise both events (i.e., one-off issues which have an impact on the organisation) and trends (i.e. time series variables which are subject to gradual change over the planning period).

The approach is designed to handle relatively large data sets. The input comprises carefully formulated subjective judgements from problem participants on probabilities of occurrence, trend movements, and pairwise interaction effects.

A time-dependent simulation model is used to generate a consistent set of the likely scenarios; consistent in that they are mutually exclusive and their probabilities of occurrence sum to one. This provides a basic framework for decision analysis and adaptive planning.

An application to an Australian state railway is then considered. The findings provide a series of detailed pictures of likely futures with which the organisation may be confronted in the next five years.

1. INTRODUCTION

Scenarios are becoming an increasingly popular tool for planning and evaluation within both public and private sector organisations. Typically, the construction of scenarios, describing possible operating environments for a corporation in the future, involves a number of basic steps.

1.1 Determining events or developments that might occur within a planning time frame and that would result in significant conseqeunces - either positive or negative - for the company;

1.2 Assigning or estimating a range of values for each of the key variables selected;

1.3 Determining the resulting interactions between these variables; and

1.4 Developing descriptions of the future under the various sets of operating conditions.

In this paper a model is outlined to generate a structured uncertainty context, i.e. to perform tasks 2, 3 and 4. The identification of the issue set (task 1) has been discussed extensively, within a particular health care problem context. [14]

Numerous alternative approaches have been suggested for undertaking scenario building exercises. Techniques such as structural modelling [6] and cross-impact analysis [2],[4], [9], [1] have been used to provide decision makers with a structured, rigorously defined context in which to assess alternative proposals or to develop new strategies. Scenario generation techniques, reviewed in [10], may be thought of as an extension of the cross-impact concept, the aim being to provide a variety of alternative likely/unlikely, desirable/undesirable, important or significant futures for input to the planning process.

2. SCENARIOS AND CORPORATE PLANNING

Corporate planning with scenarios is one method of dealing with the unpredictability of the future. Although the intention is to provide insight into possible changes and developments, the aim is not to predict, in a statistical sense, the future. Scenarios are intended to raise awareness and to avoid surprise. To this end they attempt to incorporate advances in technology, changes in social attitudes, changes in legal and administrative systems, possible

modifications to international agreement and future government priorities and any other developments which are, to a large degree beyond the control of the decision maker.

Scenarios must be described in fairly concrete terms to enable evaluation and systematic study of the future. The variables or elements that are currently included within most scenarios may be considered in two broad classes namely, events and trends.

An event is defined as a binary phenomenon that may occur either once and only once or else not occur in a specified time frame. Events are characterised by occurring suddenly and having an impact on the performance of the system. That impact may be significant or insignificant, desirable or undesirable. As the objective is to provide the decision maker with a rich picture, events may include social, political, technological, economic, legal, and behavioural changes.

Trends, on the other hand, are movements in time series data that ordinarily change gradually with time (e.g., Comsumer Price Index). Unlike events, trends do not either "occur" or "not occur", but from time to time there are sudden and unexpected deviations or perceived movements that may cause some degree of consternation.

A trend statement is also an alternative to a complex set of specific event statements and, in this sense, provides a valuable tool for including issues that are vague and ill-defined yet potentially of importance. For example, rather than attempting to specify a number of precise levels of unemployment for particular years, it is possible to consider likely deviations about a known or postulated trend.

To be useful, most scenario generation procedures must include both events and trends and some sense of their possible interaction. Events and trends provide the raw material "blocks" for the scenarios. The art of construction is in shaping or rearranging these blocks and focussing on the most important of the interactions.

Ways of including scenarios within the planning process [5], [12], [15] range from a heuristic, unstructured discussion of uncertainty within the realm of the planning process to a formal decision making under uncertainty multiple-objective-trade-off algorithm. The model outlined in this paper may be used either to provide general insights into the corporate planning process, as a framework in which

to evaluate alternative options or as a formal input to an adaptive planning process. One significant feature of this model is its ability to handle a relatively large number of issues. Whilst the principle of parsimony is taken as axiomatic for model building the same is not true when attempting to structure a rich and varied picture of the environment.

A final characteristic of this approach is that the basic input comprises structured subjective judgements of participants in the problem process. Methods of eliciting subjective data and of combining expert judgements have been discussed elsewhere [17], [11]. This paper is concerned with the task of structuring such data for use within an analytic context.

3. SIMULATION MODEL FOR SCENARIO GENERATION

The basic simulation model for scenario generation has five general stages:

3.1 Specify Input

3.2 Compute impact factors

3.3 Perform simulation

3.4 Select highest probability scenario and adjust parameters

Steps 3.2,3.3 and 3.4 are repeated until the desired number of scenarios are generated (i.e. a consistent set of the likely scenarios).

Each stage in the process is now outlined in some detail. Data is restricted to pairwise probability relationships as there are no available practical procedures for subjectively eliciting third and higher order information.

3.1 Specify Input

Consider the case in which there are n events and m trends. Data is required on trend levels, event probabilities and all interactions. The input may be partitioned as follows: Event x Event, Trend x Trend, Event x Trend and Trend x Event. Each partition is discussed in turn and brief comment is offered on the procedures used to estimate the various parameters.

3.1.1 Event x Event

This comprises probability and joint probability information on events.

For each event E_i, $i = 1, \cdots, n$,

Let $P(E_i) = P_i$ = probability of occurrence of event i within the specified time period and

$P(E_i \cap E_j) = P_{ij}$ = joint probability of occurrence of both events (i,j) within the time period,

where event probabilities are computed by averaging a set of subjective estimates of the likelihood of occurrence from a sample of individuals.

The explicit inclusion of time complicates the computation of the joint probabilities as the usual classical statistical procedures are non-time dependent. Some approaches use cumulative distribution functions to improve event estimates but the estimation of interactions, unless itself a dynamic process, suffers all the limitations of this approach [13], [14].

The issue of determining the relationship between time dependent conditional probabilities and traditional conditional probabilities has been discussed extensively [3] and the exact relationship has been demonstrated. In this context, an approximation formula [13] may be used to estimate P_{ij}, viz,

$$P_{ij} \doteq P_{i/j_f} \cdot P_{j_{f/i \cap j}} \cdot P_j + P_{j/i_f} \cdot P_{i_{f/i \cap j}} \cdot P_i$$

where $P_{i/jf}$ = probability of event i given that event j has occurred (first), and

$P_{j_{f/i \cap j}}$ = probability that event j occurs first given that both events i and j occur in the time period.

These probabilities may now be estimated directly.

3.1.2 Trend x Trend

This comprises probability distributions of trend levels

and average trend values.

For each trend T_j, $j = 1, \ldots m$,

Let $P(T_j = S_k) = f_{j(k)}$

where S_k = the k^{th} state of the discrete trend distribution, $k = 1, \ldots, r$

$f_{j(k)}$ = probability that trend j is in state k and

$$E(T_j) = \sum_{k=1}^{r} f_{j(k)}\ v(S_k)$$

= expected value of trend j,

where $v(S_k)$ = value attached to state S_k of trend j and

$$E(T_j^2) = \sum_{k=1}^{r} f_{j(k)}\ v^2(S_k)$$

= second moment

For each pair of trends T_i, T_j $i \neq j$, a revised probability distribution is calculated, namely;

$$P(T_j = S_k \;/\; T_i = S_{k'}) = f_{j(k)/i(k')}$$

where $k' = 1, \ldots, k, \ldots, r$.

Computation:

Trends are treated as discrete variables with a range of possible states around a "no change" state. Participants provide values of the most likely state for the trend at the end of the specified time period and a probability distribution, an expected value and a second moment are computed for each trend.

The computation of the "trend on trend impact" is a time consuming exercise empirically if an attempt is made to estimate a conditional trend distribution $(f_{j(k)/i(k')})$ for every value of k'. An alternative approach is to estimate the distribution for trend j say given the highest value of T_i $(v(S_r))$ and the lowest value of $T_i(v(S_1))$ and to interpolate for intermediate states,

For example, the interpolated conditional trend state

probability for trend j is given that trend i is in state t (S_t), $1<t<r$, is

$f^*_{j(k)/i(t)}$ $(1 \leq t \leq r)$, and is given by

$$f^*_{j(k)/i(t)} = \begin{cases} f_{j(k)} + \dfrac{\left(f_{j(k)/i(r)} - f_{j(k)}\right)\cdot\left(v(S_t)-E(T_j)\right)}{v(S_r) - E(T_j)} & \text{if } v(S_t)>E(T_j) \\ f_{j(k)} - \dfrac{\left(f_{j(k)} - f_{j(k)/i(1)}\right)\cdot\left(E(T_j)-v(S_t)\right)}{E(T_j) - v(S_1)} & \text{otherwise} \end{cases}$$

The actual requirement of respondents can either be to specify a new (conditional) probability distribution or to extend the magnitude of the change (if any) in the original distribution.

3.1.3 Event x Trend

This comprises revised probability distributions of trend levels given the occurrence of an event ($E_i=1$) and non-occurrence of an event ($E_i=0$) respectively.

For each trend T_j, $j=1, \ldots, m$, and
for each event E_i, $i=1, \ldots n$

let, $P(T_j=S_k/E_i=1) = f_{j(k)/i}$ and

$P(T_j=S_k/E_i=0) = f_{j(k)/\bar{i}}$

Computation:

These probability distributions are estimated similarly to those in the trend x trend section, except that there is now no need for interpolation.

3.1.4 Trend x Event

This comprises revised probability extimates of events given specific levels of trends.

For each event E_i, and trend, T_j,

let, $P(E_i/T_j = S_k) = P_{i/j(k)}$ $k=1, \ldots r$

= probability of occurrence of E_i given that trend T_j was in state S_k

Computation:

The revised probability estimates are averages of a set of subjective estimates of the conditional outcomes. For practical pruposes only, upper and lower state values of the trend are used and intermediary values are computed by interpolation. For example,

let, $S_k = t$, $1<t<r$, for trend T_j

then,

$$P^*_{i/j(t)} = \begin{cases} P_i + \dfrac{\left(P_{i/j(r)} - P_i\right)\cdot\left(v(S_t) - E(T_j)\right)}{v(S_r) - E(T_j)}, & \text{if } v(S_t)\ E(T_j) \\[2ex] P_i - \dfrac{\left(P_i - P_{i/j(1)}\right)\cdot\left((E(T_j) - v(S_t)\right)}{E(T_j) - v(S_1)}, & \text{otherwise} \end{cases}$$

3.2 Compute Impact Factors

As stated earlier, the probability measures are used to construct a set of impact factors which in turn form the basic structure of the simulation. These factors indicate the magnitude of change to a trend level or event probability. The first round effects are simply deduced from the (event, event), (trend,event) and (trend, trend) conditional probabilities outlined above. Once the original values have been changed, however, we require rules to adjust probabilities given the occurrence (or non-occurrence) of events and/or the changes in the states of the various trends. These rules are reflected in a set of impact factors.

3.2.1 Event x Event

Let $R_{j/i}$ = factor measuring impact that the occurrence of event E_i has on E_j and

$R_{j/\bar{i}}$ = factor measuring impact that non-occurrence of event E_i has on E_j.

Then,

$$R_{j/i} = \begin{cases} \left(\dfrac{P_{ij}}{P_i} - P_j\right)/(1-P_j), & \text{if } \dfrac{P_{ij}}{P_i} > P_i \\[2ex] \left(\dfrac{P_{ij}}{P_i} - P_j\right)/P_j, & \text{otherwise}. \end{cases}$$

$$R_{j/\bar{i}} = \begin{cases} \left(\dfrac{P_j - P_{ij}}{1 - P_i} - P_j\right) / (1-P_j), & \text{if } \dfrac{P_j - P_{ij}}{1 - P_i} > P_j \\ \left(\dfrac{P_j - P_{ij}}{1 - P_i} - P_j\right) / P_j, & \text{otherwise.} \end{cases}$$

3.2.2 Trend x Trend

Let $R_{j(k)/i(t)}$ = impact factor that trend T_i in state S_t has on trend T_j in state S_k

Then,

$$R_{j(h)/i(t)} = \begin{cases} \dfrac{f^*_{j(k)/i(t)} - f_{j(k)}}{1 - f_{j(k)}}, & \text{if } f^*_{j(k)/i(t)} > f_{j(k)} \\ \dfrac{f^*_{j(k)/i(t)} - f_{j(k)}}{f_{j(k)}}, & \text{otherwise.} \end{cases}$$

3.2.3 Event x Trend

Let, $R_{j(k)/i}$ = impact factor that the occurrence of event E_i has on trend T_j in state S_k

$R_{j(k)/\bar{i}}$ = impact factor that the non-occurrence of event E_i has on trend T_j in state S_k

Then,

$$R_{j(k)/i} = \begin{cases} \dfrac{f_{j(k)/i} - f_{j(k)}}{1 - f_{j(k)}}, & \text{if } f_{j(k)/i} > f_{j(k)} \\ \dfrac{f_{j(k)/i} - f_{j(k)}}{f_{j(k)}}, & \text{otherwise, and} \end{cases}$$

$$R_{j(k)/\bar{i}} = \begin{cases} \dfrac{f_{j(k)/\bar{i}} - f_{j(k)}}{1 - f_{j(k)}}, & \text{if } f_{j(k)/\bar{i}} > f_{j(k)} \\[2ex] \dfrac{f_{j(k)/\bar{i}} - f_{j(k)}}{f_{j(k)}}, & \text{otherwise.} \end{cases}$$

3.2.4 Trend x Event

Let, $R_{i/j(t)}$ = impact factor of trend T_j in state t (S_t) on event i

Then,

$$R_{i/j(t)} = \begin{cases} \dfrac{P^*_{i/j(t)} - P_i}{1 - P_i}, & \text{if } P^*_{i/j(t)} > P_i \\[2ex] \dfrac{P^*_{i/j(t)} - P_i}{P_i}, & \text{otherwise.} \end{cases}$$

3.3 Perform Simulation

The aim of each simulation is to produce, for n events and m trends, estimates of the probability of occurrence of each of the 2^n event-scenarios together with the corresponding average values of each of the m trends. This is achieved by generating N sample scenarios based on the event and trend probability information discussed above. Each sample requires (n + m) passes at which one of the scenario elements is determined and the event and trend probabilities for undetermined elements are adjusted. As explained in section 3.4 this whole simulation procedure is repeated with modified input probabilities and impact factors and from each such iteration the highest probability event-scenario is "removed" from the probability space.

The basic simulation process is outlined below in steps (i) through (ix):

(i) Select an undetermined event (E_i) or a trend (T_j) at random.

If event (E_i) is selected in (i), do (ii) and (iii).

(ii) Determine occurrence or non-occurrence using random numbers and the current probability value.

(iii) Compute for each event, E_j ($j \neq i$) refined probabilities, P_j^* and for each trend T_j, refined or modified trend frequency distributions $f^*_{j(k)}$, $k=1, \ldots, r$
where, for occurrence of E_i:

$$P_j^* = \begin{cases} P_j + R_{j/i}(1-P_j) & \text{, if } R_{j/i} > 0 \\ P_j\,(1+R_{j/i}) & \text{, otherwise , and} \end{cases}$$

$$f^*_{j(k)} = \begin{cases} f_{j(k)} + R_{j(k)/i}\,(1-f_{j(k)}) \text{ ,} & \text{if } R_{j(k)/i} > 0 \\ f_{j(k)}\,(1+R_{j(k)/i}) \text{, otherwise.} & \end{cases}$$

and, for non-occurrence of E_i:

$$P_j^* = \begin{cases} P_j + R_{j/\bar{i}}\,(1-P_j) \text{ , if } R_{j/\bar{i}} > 0 \\ P_j\,(1+R_{j/\bar{i}}) \text{ , otherwise , and} \end{cases}$$

$$f^*_{j(k)} = \begin{cases} f_{j(k)} + R_{j(k)/\bar{i}}\,(1-f_{j(k)}) \text{ ,} & \text{if } R_{j(k)/\bar{i}} > 0 \\ f_{j(k)}(1+R_{j(k)/\bar{i}}) \text{, otherwise.} & \end{cases}$$

The revised values $f^*_{j(k)}$ derived above (and in (v) below) are also rescaled to ensure that $\sum_{k=1}^{r} f^*_{j(k)} = 1$

On the second and subsequent passes the "current" probabilities and trend state probability values $f_{j(k)}$ will themselves have been modified by previous event and/or trend determinations. This will also apply to the trend computations below.

If trend T_i is selected in (i), do (iv) and (v):

(iv) Using random numbers, determine the trend state, S_t, and value $v(S_t)$ using the current frequency distribution for trend T_i, $f_{i(k)}$, $k = 1, \ldots r$.

(v) Compute for each event, modified probabilities P_j^* and for each other trend $T_j (j \neq i)$, modified trend frequency distributions $f_{j(k)}^*$, $k = 1, \ldots r$, as follows:

$$P_j^* = \begin{cases} P_j + R_{i/j(t)}(1-P_j) & \text{, if } R_{i/j(t)}>0 \\ P_j\left(1 + R_{i/j(t)}\right) & \text{, otherwise, and} \end{cases}$$

$$f_{j(k)}^* = \begin{cases} f_{j(k)} + R_{j(k)/i(t)}\left(1 - f_{j(k)}\right), & \text{if } R_{j(k)/i(t)}>0 \\ f_{j(k)}\left(1 + R_{j(k)/i(t)}\right), & \text{otherwise} \end{cases}$$

(vi) Repeat step (i) and either (ii) and (iii), or (iv) and (v) until all events and trends have been determined, i.e. (a) all events have either occurred or not occurred and (b) all trends have been assigned a trend value. This determines one simulated sample scenario and requires n + m passes.

(vii) Record: (a) The cumulative frequency of each of the 2^n event-scenarios.
(b) For each event-scenarios, the cumulative sum of the trend values and their squares for all trends.

Thus, after N samples we would have accumulated:

g_l = number of occurrences of event-scenario l, $l = 1, \ldots, 2^n$

$TT_{jl} = \sum_{h=1}^{g_l} v(S_{t,h})_l$ = sum of the values of Trend T_j which have occurred with event-scenario l

$$TT_{jl} = \sum_{h=1}^{g_l} v^2(S_{t,h})_l = \text{sum of squares of the values of Trend } T_j \text{ which have occurred with event-scenario } l$$

where $v(S_{t,h})_l$ is the value associated with the state S_t determined as the state of trend, T_j occurring with the hth occurrence of event-scenario l.

(viii) Calculate, for each of the 2^n event-scenarios

$Q_l = \frac{g_l}{N}.G$ = estimate of the probability of occurrence of event-scenario $l, l = 1, \ldots, 2^n$

where G = reduced probability space at this iteration

$\bar{T}_{jl} = \frac{T_{jl}}{g_l}$ and $S_{jl} = \sqrt{\left(\frac{TT_{jl}}{g_l} - \bar{T}^2_{jl}\right)}$, the mean

and standard deviation, respectively, of Trend T_j associated with event-scenario l,

$j = 1, \ldots m$, and $l = 1, \ldots, 2^n$

(ix) Rank the event-scenarios on Q_l.

3.4 Selection of Scenarios and adjustment of Parameters

The basic approach is to select one of the event-scenarios from each simulation as being suitable, "remove" it from the probability space being studied, adjust the parameters accordingly and to repeat the simulation process until sufficient suitable scenarios have been selected. Ideally, we seek to remove the most likely scenario, i.e. the one with the highest probability, on each iteration. Such an approach guarantees a set of scenarios which span the space and which generally provide considerable variety for the decision maker [8].

The concept of the most likely scenario is difficult to define in absolute terms as the probabilities of trends moving from one state to another have not been computed. For event-only scenarios, a scenario likelihood measure can be determined by considering all the possible event outcomes in a mixed integer program formulation [8]. In our simulation approach, the Q_l values for the event-scenarios are the

obvious measure but they only implicitly include the effects of trend impacts and interactions. In some sense, the scenarios may be envisaged as being generated by first identifying a set of fundamental, or basic, trends (and their status quo extrapolations) and then by considering the overall effect of a series of disturbances (events) on these projections. The outcome is then an event set which comprise a series of occurrences and non-occurrences, and a resultant set of average trend values, $\bar{T}_{jl}$. The trends themselves, could possibly be partitioned into those which are critical indicators of environmental change and those which are important, either in their own right or because of an interaction (or derived) effect.

Given that a scenario has been selected for removal, it is necessary to adjust the initial event probabilities, trend distributions and impact factors to reflect the reduced probability space. The general algorithms for making these adjustments are outlined below. (The formulae give adjusted values after k iterations, k = 1, ..., K)

(a) Event Probabilities

$$P^*_i = \frac{P_i - \sum_{k=1}^{K}\left(Q\ max_k \cdot \delta_{i,k}\right)}{G}$$

where $Q\ max_k$ = the probability of the maximum probability scenario at iteration, k, (i.e. the scenario removed at that iteration)

$$\delta_{i,k} = \begin{cases} 1, \text{if event, } E_i\text{, occurred in the} \\ \quad \text{scenario removed at iteration, } k \\ 0, \text{otherwise} \end{cases}$$

and

$$G = 1 - \sum_{k=1}^{K} Q\ max_k = \text{ the reduced probability space}$$

Similarly,

$$P^*_{ij} = \frac{P_{ij} - \sum_{k=1}^{K}\left(Q\ max_k \cdot \delta_{ij,k}\right)}{G}$$

$$\text{where } \delta_{ij,k} = \begin{cases} 1, \text{if events } E_i \text{ and } E_j \text{ both occurred in the scenario removed at iteration } k \\ 0, \text{otherwise} \end{cases}$$

The division by the remaining probability space, G, in the above relationships provides the subsequent iterations with an apparently full probability space. Subsequent values of event scenario probabilities are reduced by the same factor (see step (viii) in section 3.3).

(b) Trend Probabilities

The scenario removed at each iteration has an associated average value of each trend T_j, $j=1, \ldots m$. The expected value of each trend will thus be modified to

$$E^*(T_j) = \frac{E(T_j) - \sum_{k=1}^{K} \bar{T}_{jl} \cdot Q \max_k}{G}$$

where $\bar{T}_{jl}$ is the average value of Trend T_j associated with the event-scenario removed at iteration, k.

This new value implies a modified set of state probabilities for each trend, such that

$$E^*(T_j) = \sum_{k=1}^{r} f^*_{j(k)} v(S_k)$$

This relationship does not uniquely determine the $f^*_{j(k)}$ probabilities but imposes one constraint on the values. Another constraint could be imposed by the second moment relationships, but we have found the following adjustment procedure quite suitable:

(i) Determine whether $E^*(T_j)$ is greater (or less) than $E(T_j)$

(ii) Determine the common proportion (β) of each state probability which must be allocated to the next higher (or lower) state, in order to achieve the new expected value, except for the highest (or lowest) state.

(iii) If $\beta \geq 1$, move all state probabilities by one state, accumulating the highest (or lowest) state and making the lowest (or highest) state probability zero. Recalculate a new $E(T_j)$ from these new probabilities and repeat from (i).

(iv) If $\beta < 1$, adjust the state probabilities, as follows, for example if $E^*(T_j) > E(T_j)$:

$$\begin{cases} f^*_{j(1)} = (1-\beta) f_{j(1)} \\ f^*_{j(k)} = (1-\beta) f_{j(k)} + \beta f_{j(k-1)}, k=2, \ldots, r-1 \\ f^*_{j(r)} = \qquad f_{j(r)} + \beta f_{j(r-1)} \end{cases}$$

$$\text{where} \quad \beta = \frac{E^*(T_j) - E(T_j)}{v(S_r) f_{j(r-1)} + \sum_{k=2}^{r-1} v(S_k) \cdot \left(f_{j(k-1)} - f_{j(k)} \right) - v(S_1) f_{j(1)}}$$

The above procedure is not unique but does appear to give satisfactory results, allowing the expected value of the trend to be adjusted to reflect the removal of an event-scenario from the probability space.

(c) Impact Factors

Event x event impact factors are recalculated using the revised values of P_i and P_{ij} given in (a) above. Although there may be a case for adjusting the interaction effects involving trends, we have preferred at this stage, to leave all the impact factors unchanged and allow them to operate on the revised event and trend probabilities derived above.

In summary, the simulation procedure outlined in 3.3 is repeated using the revised probabilities and impact factors given in this section. Following each simulation, the highest probability event-scenario is removed and the probabilities revised accordingly. The whole procedure is repeated until the desired number of (removed) scenarios has been obtained. These scenarios, which should be reasonably likely and diverse, provide a basic input for corporate planning and decision making.

CASE STUDY - Corporate Scenarios for Westrail

This iterative simulation model has been applied as part of an overall approach to scenario generation, to generate a set of scenarios for use as part of the annual planning cycle for the Western Australian Government Railways (Westrail). Westrail operates a rolling five year plan and hence attention was focussed on those aspects of the environment in which the railway will operate which might change or which might require action in the next five years. The corporate profit model was then evaluated for a number of development options in the light of the scenarios generated.

The overall approach is detailed in [10] and involves the following steps:

1. Search the environment for items (events and trends) which are relevant to the problem context.
2. Assess the items.
3. Search for related groups of items.
4. Assess the cross-impact between items within each group.
5. Use some scenario generation procedure for each group.
6. Give each generated scenario a descriptive title and repeat steps 2-4 with these "super-items".
7. Generate overall scenarios.

For the Westrail study, the procedure used in step 5 on each related group of items was, generally, that described in this paper.

The participants in the study, whose judgements were the prime material and determinants of the final scenarios, comprised over 30 senior executives from all areas of the organisation.

In group sessions, these participants generated 96 items in the form of 53 events and 43 trends. These items were assessed by the group using the fixed time-period model outlined. An analysis of these assessments was distributed to participants. This included histograms of response frequencies, a graph of average probability against average desirability for events and graph of average movement against desirable movement for trends, together with a verbal summary of the results.

At this stage a number of items, were dropped from the next phase of the analysis. Items were excluded if they were perceived after assessment to have little significance for the organisation. Some events were assessed as being very unlikely or almost certain and were thus not included in the

sub-group generation stages. (They were included in the final statements of the overall scenarios as having not occurred or occurred, respectively).

The identification of groups of related items within the reduced set of 68 items was achieved by the analysts, mainly on the basis of an imposed structure suggested by the items themselves. This analysis revealed eight groups of items which appeared to have significant internal interrelationships. A number of events and trends appeared in more than one group. Each group represented an aspect of the environment and was given a description. (Table 1). Some groups contained 16 or 17 items and a description of all items and scenarios may be found in [10].

A full cross-impact analysis was conducted for each group. The interaction between all pairs of items was studied, using the procedures detailed in this paper. One result of structuring into eight areas was to reduce the number of pairs of items to be examined from over 2300 to about 600. This was still a large task as each pair required two or three questions to be answered. By seeking only five individual opinions on each pair the thirty odd participants were still asked to consider nearly 100 pairs.

Scenario generation within each group was performed by a mixture of simulation and mixed integer/linear programming. The full iterative scheme described in 3.4 above was approximated by using the event-only data to generate the most likely scenario for removal at each iteration. This technique is described in [8]. The first two or three event-scenarios selected, together with the simulated average trend values corresponding to those event combinations were used as the representative scenarios for each group of items. These were given descriptive titles which are reported in Table 1. An example is shown in Table 2 for the Operations area of Westrail's environment (Group G).

Some of the larger groupings (i.e. 16 or 17 items) were only computationally tractable in reasonable time because of the existence of some sets of mutually exclusive events. In retrospect, it would certainly have been preferable to further structure the original issue set to reduce the number of issues in each group.

Five overall scenarios were then produced by considering the possible combinations of the alternative scenarios in the eight areas. This synthesis was aided by the existence of common events and trends, and some obvious pairings which

these common elements indicated. Table 3 shows the final scenarios with the titles derived for easy reference and identification. These scenarios were discussed by the participants and their implications for Westrail were studied. The scenarios were then used within the organisation's existing annual planning cycle. The five overall scenarios were reduced to three by analysts in Westrail's Management Services Bureau. These formed part of the input considered by Westrail's Strategic Planners at their meeting to initiate the planning cycle for the coming year.

CONCLUSION

Scenarios, if carefully formulated, provide a comprehensive framework is which planning and evaluation of strategies may take place.

The model outlined in this paper provides a realistic step towards providing management with a comprehensive set of likely scenarios. Some extensions to this model are obvious. The inclusion of multiple time periods may improve the usefulness of the scenarios. The development of the set of programs in interactive mode may further enhance the applicability of the technique. Finally, the inclusion of options, i.e. different algorithms, at various stages in the process enables the planner to best utilise his own resources.

Scenario building, at best, is a "means-oriented" exercise. The ultimate "end" is to use the scenario, for whatever purpose, within the adaptive or flexible planning process.

Table 1

SCENARIO SUB-GROUPS

Group	Area	Titles of Selected Scenarios	
A	Industrial Relations	AI	Industrial Peace
		AII	Unions Down - But Not Out
		AIII	Industrial Trouble
B	Social Environment	BI	Increased Leisure
		BII	The Work Ethic Continues
C	Government Decisions Affecting the Railway's Operations	CI	Transport Policy Rationalised
		CII	Rail Restricted but Protected
		CIII	Declining Railway Influence
D	Government Decision Affecting the Railway's Financial Arrangements	DI	Rational, Planned, Efficient Competition
		DII	Declining Railway Influence
		DIII	Restricted Competition - Rail Protected
E	Demand	EI	Mineral Expansion
		EII	Overall Expansion - New Developments
		EIII	No Major Growth
F	International/Energy	FI	"Steady" Economic Situation
		FII	Local Oil Crisis
		FIII	"Insulated Energy Cost" Situation
F	Operations	GI	Technological Development
		GII	Some Development at a Price
		GIII	No Development - Industrial Trade-off
H	Other Transport Modes (Competition)	HI	Open Free Competition
		HII	Restricted Competition
		HIII	Green Light for Road Transport

Table 2

GROUP 2: OPERATIONS SCENARIOS

SCENARIO ELEMENTS

Events		Probability
20.	Union opposition forces abandonment of use of contract labour.	.42
22.	Ban by grain handling authority staff on unloading other than rapid discharge wagons.	.37
23.	Union permits two man train crews.	.39
28.	Effective techniques for bulk fertiliser handling introduced.	.55
31.	Use of lighter metal increases the load/tare ratio.	.45

Trends		Trend Movement
22.	Efficiency of diesel engines.	.33
23.	Complexity of equipment.	.54
24.	Equipment obsolescence rate.	.50
31.	Westrail's traffic task for grain.	.49
37.	Supply of skilled labour for Westrail.	.07
38.	Cost of spare parts in real terms.	.67

LIKELY SCENARIOS

	Events					Trends					
No.	20	22	23	28	31	22	23	24	31	37	38
I	0	0	0	1	1	.36	.42	.21	.51	.22	.58
II	0	1	1	1	0	.37	.42	.34	.39	.06	.62
III	1	0	0	0	0	.42	-.19	.08	.54	.19	.60

I Technological Development

II Some Development at a Price

III No Development - Industrial Trade-Off

Table 3

OVERALL SCENARIOS

SCENARIO 1: *Opportunity, development and growth with minimal difficulties*

AI	Industrial Peace	BI	Increased leisure
CI	Transport policy rationalised	DI	Rational, planned, efficient competition
EII	Overall expansion - new developments	FII	Local oil crisis
GII	Some (technological) developments at a price	HI	Open free competition (with other modes)

SCENARIO 2: *Steady progress and growth with some problems*

AII	Unions down - but not out	BII	The work ethic continues
CI	Transport policy rationalised	DI	Rational, planned, efficient competition
EI	Mineral expansion	FII	Local oil crisis
GI	Technological development	HI	Open free competition (with other modes)

SCENARIO 3: *Opportunities frustrated by lack of growth and industrial trouble*

AIII	Industrial trouble	BII	The work ethic continues
CI	Transport policy rationalised	DI	Rational, planned, efficient competition
EIII	No major growth	FII	Local oil crisis
GIII	No (technological) development	HI	Open free competition (with other modes)

SCENARIO 4: *Restricted and controlled steady progress*

AII	Unions down - but not out	BII	The work ethic continues
CII	Rail restricted but protected	DIII	Restricted competition - rail protected
EI	Mineral expansion	FII	Local oil crisis
GI	Technological development	HII	Restricted competition

SCENARIO 5: *Hard times - railway in decline*

AII	Unions down - but not out	BII	The work ethic continues
CIII	Declining railway influence	DII	Declining railway influence
EIII	No major growth	FI	Steady economic situation
GI	Technological development	HIII	Green light for road transport

REFERENCES

[1] Alter, S., THE COMPUTATIONAL MATHEMATICS OF TIME-DEPENDENT CROSS-IMPACT MODELLING, Centre for Futures Research, M26, USC, 1976.

[2] Enzer, S., CROSS-IMPACT TECHNIQUES IN TECHNOLOGY ASSESSMENT, Futures, Vol 4, No 1, PP. 30-51, 1972.

[3] Enzer, S. and S. Alter, CROSS-IMPACT ANALYSIS AND CLASSICAL PROBABILITY, Futures, Vol 10, NO 3, 1978.

[4] Kane, J., A PRIMER FOR A NEW CROSS-IMPACT LANGUAGE - KSIM, Technological Forecasting and Social Change, 4, PP. 129-142, 1972.

[5] Linneman, R.E. and J.D. Kennell, SHIRT-SLEEVE APPROACH TO LONG RANGE PLANS, Harvard Business Review, March-April, 1977.

[6] Linstone, H.A. et al, THE USE OF STRUCTURAL MODELLING FOR TECHNOLOGY ASSESSMENT, vols I and II, Report to the National Science Foundation, Portland State University, February 1978.

[7] Lipinski, H. and J. Tydeman, CROSS-IMPACT ANALYSIS - MODIFIED KSIM, Submitted to Futures, 1978.

[8] Mitchell, R.B., J. Tydeman and R. Curnow, SCENARIO GENERATION : LIMITATIONS AND DEVELOPMENTS IN CROSS-IMPACT ANALYSIS, Futures, Vol 9, No 3, PP. 205-215, 1977.

[9] Mitchell, R.B. and J. Tydeman, SUBJECTIVE CONDITIONAL PROBABILITY MODELLING, Technological Forecasting and Social Change, 11, PP. 133-152, 1978.

[10] Mitchell, R.B., J. Tydeman and J. Georgiades, STRUCTURING THE FUTURE - THE APPLICATION OF A SCENARIO GENERATION PROCEDURE, Technological Forecasting and Social Change, forthcoming 1979.

[11] Morris, P.A., DECISION ANALYSIS EXPERT USE, Mgt Sci, Vol 20, No 9, 1974.

[12] O'Connor, R., PLANNING UNDER UNCERTAINTY : MULTIPLE SCENARIOS AND CONTINGENCY PLANNING, Research Report, Conference Board Canada, 1978.

[13] Tydeman J. and R.B. Mitchell, SUBJECTIVE INFORMATION MODELLING, Opl. Res. Q'tly, Vol 28, No 1, PP. 1-19, 1977

[14] Tydeman J. and R.B. Mitchell, AUSTRALIAN HEALTH CARE FUTURES, Administrative Studies Program, Research Monograph 5, Australian National University, 1978.

[15] Tydeman J. and R.B. Mitchell, POLICY EVALUATION UNDER UNCERTAINTY : AN APPROACH USING SUBJECTIVE INFORMATION, Socio-Economic Planning Sciences, Vol 12, 1978.

[16] Wakeland, W., QSIM 2: A LOW BUDGET HEURISTIC APPROACH TO MODELLING AND FORECASTING, Technological Forecasting and Social Change, 9, PP. 213-229, 1976.

[17] Winkler, R.L., THE CONCENSUS OF SUBJECTIVE PROBABILITY DISTRIBUTIONS, Mgt Sci., Vol 15, No 2, 1978.

COMPARISON OF LOT-SIZING TECHNIQUES FOR MULTI-LEVEL MATERIAL REQUIREMENTS PLANNING SYSTEMS

EUI H. PARK

Division of Data Base Management,
Boeing Company, Seattle,
Washington 98124, U.S.A.

SUNG H. PARK

Department of Computer Science
and Statistics,
Seoul National University,
Seoul 151, Korea

ABSTRACT. This paper compares and analyzes five lot-sizing techniques (Economic Order Quantity(EOQ), Period Order Quantity(POQ), Lot-for-lot(LFL), Part Period Balancing(PPB) and Least Total Cost(LTC)) in the combined problem of shop scheduling and lot-sizing in multi-level material requirements planning systems. A brief description for each technique is given and a computer program is also developed for each technique. Simulated problems are solved by the computer programs to compare and analyze different situations which reflect typical conditions in manufacturing environments. The results obtained are statistically analyzed using the techniques of least significant difference and regression analysis.

It is the hope of the authors that this paper will assist in developing a proper lot-sizing technique for a manufacturing company under a typical manufacturing environment, and in developing a supply system of munitions for a military

logistic system.

1. INTRODUCTION

There are basically two alternatives in fundamental approach that a manufacturing enterprise may employ for purposes of inventory management. They are:

a) Order point systems, popularly known as statistical inventory control,

b) Material requirements planning.

The first approach is a set of procedures and decision rules intended to ensure continuous physical availability of all items comprising an inventory in the face of uncertain demand. The order point is determined for each inventory separately, based on the forecast demand and on the probability of actual demand exceeding the forecast.

A material requirements planning(MRP) system consists of a set of logically related proedures, decision rules, and records designed to translate a master production schedule into time-phased net requirements, and the planned coverage of such requirements, for each component inventory item needed to implement this schedule. An MRP system replans net requirements and coverage as a result of changes in either the master production schedule, or inventory status, of product composition.

Order point is part-based, whereas MRP is product-oriented. MRP systems are receiving increased attention by production and purchasing managers as many firms turn to the computer for assistance in planning and controlling manufacturing operations. These systems reduce a complicated master schedule of finished products to a time-phased schedule of requirements for intermidiate assemblies and component parts.

MRP systems include a procedure for determining the lot size and timing of replenishment orders to meet the forecast requirements. Thus, one problem often encountered in designing such a system is that of selecting a procedure for making lot size decisions. Although a number of procedures have been proposed, ranging from the use of simple decision rules to extensive optimizing procedures, there is surprisingly little guidance

for the manager in selecting a lot sizing procedure for his system.

In recent years, the traditional interest in the classical economic order quantity(EOQ) that is based on the assumption of continuous and steady-rate demand, has shifted to lot sizing in an environment of discrete period-demands. A number of lot sizing techniques for MRP systems have been proposed. Some of the most widely recognized approaches are

a) Economic Order Quantity(EOQ)
b) Lot-for-Lot(LFL)
c) Period Order Quantity(POQ)
d) Part-Period Balancing(PPB)
e) Least Total Cost(LTC)

Besides these, there is a precise mathematical technique that is based on the concept of dynamic programming, but it is hardly used in practice due to its complexity. This technique is called the Wagner-Whitin algorithm [8].

The purpose of this paper is to present a framework for comparing the performance of the five procedures mentioned above over a broad range of cost and demand data parameters. For the readers who are not familiar with the lot sizing techniques, the data in Table 1 will be used to explain the procedures.

Week Number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	50	80	180	80	0	0	180	150	10	100	95	180

Table 1: Weekly requirements schedule

The example shown in Table 1 illustrates a typical requirements forecast that is considered in planning the lot size and timing of replenishment orders. We shall only consider the problem in the context of an MRP system, i.e. when the demand forecast is derived by an explosion of finished product requirements. In this case, the schedule of weekly requirements for an individual component, like the one shown in this example, is derived by exploding the scheduled production for all higher level assemblies into the necessary

component parts. The weekly requirement for each component is then obtained by accumulating its weekly usage in all higher level assemblies.

The assumptions made in this paper are as follows. First, since the component requirements are aggregated by time period for planning purposes, we assume that all of the requirements for each period must be available at the beginning of the period. Second, all of the requirements for a given period must be met and cannot be backordered. Third, since the system is operated on a periodic basis, the ordering decisions are assumed to occur at regular time intervals, weekly. Fourth, the orders which are placed at the beginning of a period, are assumed to be available in time to meet the requirements for that period. Finally, the components are withdrawn from inventory at a uniform rate during each period. Therefore, the average inventory level will be used in computing the inventory carrying costs.

2. LOT SIZING TECHNIQUES

In this chapter we shall illustrate the results obtained by applying the five procedures to the example shown in Table 1. The following cost values are used for this example.

Ordering cost(S) = \$300 per order
Inventory carrying cost(C) = \$2 per unit per week

2.1 Economic Order Quantity(EOQ)

The economic lot size formula is often used as a decision rule for placing orders in an MRP system because of its simplicity [3]. The economic order quantity is calculated as:

$$Q = \sqrt{2US/C}$$

where U is the average weekly demand. In this example U is equal to 92.1 units for the entire requirements schedule, and the EOQ is 166. Table 2 shows the sequence of orders obtained by the EOQ formula.

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	50	80	180	80	0	0	180	150	10	100	95	180
Planned-order Coverage	166	0	166	166	0	0	166	166	0	0	166	166

Table 2: Economic order quantity example

This example points out a problem with the EOQ procedure. Since the demands do not occur at a continuous and steady-rate, as is assumed by the EOQ formula, the restriction of fixed lot-sizes results in larger inventory carrying cost. This occurs because of the mismatch between the demand and the order quantities, causing excess inventory to be carried forward from week to week. In the above illustration, 36 units are carried in inventory from weeks 2 to 3 whereas these items could have been included in the order placed in week 3. Consequently, the more discontinued and non-uniform demands are, the less effective the EOQ will be.

2.2 Lot-for-Lot(LFL)

This technique, also called "discrete ordering", is the simplest and most straight forward of all lot-sizing techniques. Since it provides the quantity of each week of net requirements, the planned order quantity always equals the quantity of the net requirements being covered. Lot-sizing by LFL results in the minimum inventory carrying cost. For this reason this type of technique is used for expensive items that have discontinuous demands whether they are manufactured or purchased. The items which have high-volume production and the items which pass through specialized facilities are also ordered Lot-for-Lot. But it will often bring the highest set-up or ordering cost.

2.3 Period Order Quantity(POQ)

As discussed in Berry [1], Orlicky [4] and Theisen [7], this technique, which is based on the logic of the classical EOQ, was revised for

use in an environment of discontinuous demand periods. By dividing the EOQ by the mean demand rate, the ordering interval would be determined. The order quantity(lot-size) is decided by this frequency. In the example the time interval would be 1.8 weeks(166/92.1 = 1.8).

Since this fixed-interval technique avoids the mismatch between order quantity and demand value, it might reduce inventory carrying cost. For this reason, the POQ is more effective than the EOQ, because set-up cost per year is the same, but carrying cost will be lower than POQ. But, because the traditional EOQ is based on infinite periods, the number of orders could be different on finite periods. When the example is applied, as is shown in Table 3, it yields the same number of orders as the EOQ procedure, but with lot sizes ranging from 50 to 330 units. Consequently, the inventory carrying cost has been reduced by 33.7%, thereby improving the total cost by 19.7% (See Table 7).

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	50	80	180	80	0	0	180	150	10	100	95	180
Planned-order coverage	50	260	0	80	0	0	330	0	110	0	95	180

Table 3: Period order quantity example

2.4 Least Total Cost(LTC)

The Least Total Cost method, discussed by Gorham [2], Orlicky[4] and Theisen [7], is primarily based on the same principle as that of the classical EOQ formula, i.e., balancing ordering and carrying costs to produce the minimum total cost. However, this technique intends to reach this objective by ordering in quantities at which the set-up cost per unit and the carrying cost per unit are nearly equal, and using the cumulative requirements as the lot-size. It makes use of the ratio of ordering to carrying costs (S/C), called the "economic part-period(EPP)"

factor. The EPP is defined as that quantity of the inventory item which, if carried in inventory for one period, would result in a carrying cost equal to the cost of setup. The part-period is a convenient expression of inventory carrying cost for purposes of comparison and tradeoff; i.e., it can be said that to carry a quantity of an item in inventory for a certain period of time will "cost" X part-periods.

The LTC technique selects that order quantity at which the part-period cost most nearly equals the EPP. An example of LTC computation appears in Table 4. The quantity chosen for the first

Week	Requirements	Carried in inventory (periods)	Part-periods (cumulative)
1	50	0	0
2	80	1	80
3	180	2	440

Table 4: Computation of LTC

lot would be 130, because the 80 part-periods that it would cost most nearly approximate the EPP of 150 (S/C = 300/2 = 150). This order would cover requirements of weeks 1 and 2, and the second order of 260 would cover requirements of weeks 3 through 6. This is shown in Table 5.

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	50	80	180	80	0	0	180	150	10	100	95	180
Planned-order coverage	130	0	260	0	0	0	330	0	110	0	275	0

Table 5: Least total cost example

2.5 Part Period Balancing (PPB)

The concept of this procedure, described by Gorham [2], is also based on the economic part period(EPP). If the number of parts held in inventory are multiplied by the number of periods over which they are held, "part-period" is derived. In determining the lot-size for an order, the PPB technique employs the same logic as LTC which is equate the set-up cost(or ordering cost) and the inventory carrying cost. But the formula used in this procedure is:

$$R = R_1/2+3R_2/2+\ldots+(n-3)R_{n-1}/2+(n-1)R_n/2 \approx S$$

where R_i is the demand in period i. The values of R are:

a) Week 1: 50/2 = 25
b) Week 1 and 2: 50/2 + 3(80)/2 = 145
c) Week 1,2 and 3: 50/2+3(80)/2+5(180)/2= 595.

Since the ordering cost of $300 is exceeded in week 3, it is time to check the differences between R and ordering cost at week 2 and week 3. The calculation for the next order should start at week 3 because the difference of week 2 is smaller than week 3. The first order quantity is accumulated for the first two weeks resulting in a lot size of 130 units.

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	50	80	180	80	0	0	180	150	10	100	95	180
Planned-order coverage	130	0	180	80	0	0	180	160	0	195	0	180

Table 6: Part period balancing example

2.6 Evaluation of Lot-Sizing Procedures

It is difficult to state categorically which lot-sizing technique is the best one, because each of the lot-sizing techniques is imperfect due

to differing assumptions and other deficiencies as mentioned above. In evaluating the relative effectiveness of these techniques, there is the difficulty that the performance of each technique varies, depending on the ratio of set-up and inventory carrying cost and the net requirements data used. The performances of the techniques using the same data set based on the assumed set-up and inventory carrying cost are compared in Table 7.

Technique	Number of set-up	Set-up cost	Carrying cost	Total cost
EOQ	7	$2,100	$2,965	$5,065
LFL	10	3,000	1,105	4,105
POQ	7	2,100	1,965	4,065
LTC	5	1,500	2,285	4,085
PPB	7	2,000	1,475	3,575

Table 7: Evaluation of lot-sizing techniques

As discussed before, the EOQ and the POQ have a same set-up cost, but the inventory carrying cost of the POQ is smaller than the one of the EOQ. While the LFL has the highest set-up cost among the five techniques, the inventory carrying cost is the lowest, because of the period by period coverage of net requirements. The LTC and PPB which approximately equate set-up and inventory carrying cost also show good total costs.

2.7 Scheduling in Multi-level Requirements System

In the determination of the lot-sizes as illustrated above, no mention is made of the capacity of the plant to produce the calculated lot-size for any given week. It is possible that a good lot-sizing technique may not be good for shop scheduling purposes. Thus, lot-sizing and shop scheduling should be considered simultaneously. Shop scheduling includes a shop capacity plan for the man, machine, and overtime needed to increase or decrease the output capacity of a particular work center.

Loading and scheduling in MRP systems are important to the production plan, since it precedes the commitment to deliver a product of given quality and quantity at a specific time.

The loading function deals with the initial assignment of a task to manpower categories such as work centers or machine centers, and scheduling is concerned with the sequence, start or completion of jobs. An ideal schedule would be one which results in:

a) all production orders being completed by their due dates,
b) zero idle time on each facility,
c) no work-in-process inventory building up between facilities.

Each of these factors has an associated cost that is usually unavoidable. Even if an ideal schedule cannot be generated, it would be desirable to generate one as close to ideal as possible. In this paper, scheduling cost which will be quantified later will be considered along with set-up and inventory carrying costs.

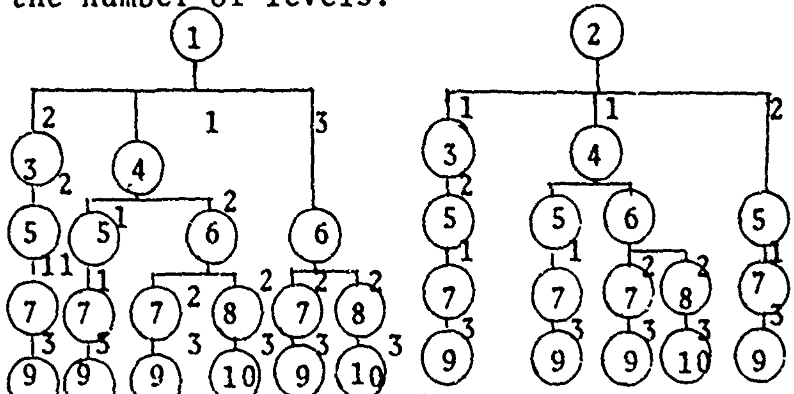
3. COMPARISON OF LOT-SIZING TECHNIQUES

In this chapter a more general framework for comparing the five different lot-sizing techniques will be presented. Berry [1] compares four techniqnes(EOQ, POQ, PPB and Wagner-Whitin Algorithm) in a single-level system where no dependencies are given through lower level requirements from components required by the end item. This situation is extended here to multi-end products with shared assemblies, sub-assemblies, and component parts.

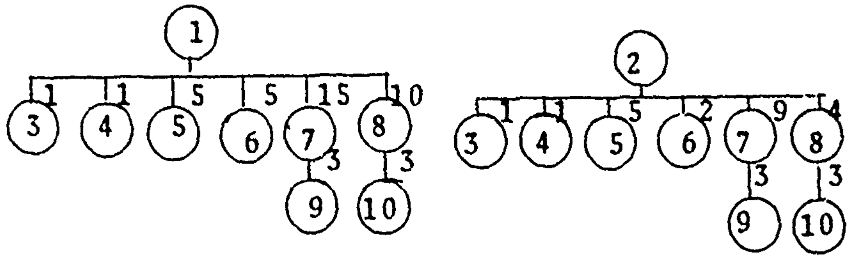
In this paper, a single lot-sizing technique is applied to all items at all levels of the product structure. There are infinitely many possible product structures that could be used to study the effects of the overall requirements in a multi-level system. For thé purpose of this research, two such product diagrams, both shown in Figure 1, will be used to illustrate the degree of component dependencies through the number of levels of assembly, and component needs. For instance, in product diagram 1, there are two end-items (1 & 2), two assembly items (4 & 6), four machining items (3,5,7 & 9) and two raw materials or components (9 & 10).

All these factors, the number of levels, the number of items, the number of end-items, the number of raw materials, the number of assembly

items, and the type of operation required, machining or assembling, were selected to reflect strong dependency relationships between items. In product diagram 2, derived from product diagram 1, the dependency requirements are reduced by lessening the number of levels.



Product diagram 1



Product diagram 2

Figure 1: Product Structure

Part numbers	1	2	3	4	5	6
Machining times			.0100		.0115	
Assembly times	.0250	.0250		.0150		.260

7	8
.0080	.0070

Table 8: Machining or assembly times for producing one unit of each part

Part numbers	1	2	3	4	5	6	7	8
Unit costs	.250	.250	.100	.150	.115	.260	.080	.070

Table 9: Unit cost of each part

Table 8 shows the machining or assembly times for producing one unit of each part, and Table 9 shows the unit cost of each part. For example, it takes 2.6 hours and costs $26.00 to produce 100 units of part number 6. In both tables, part 9 and part 10 are not included, because these parts are raw materials.

The cost factors considered to compare the lot-sizing techniques are inventory, set-up and scheduling costs. Each cost factor is influenced by the following variables: economic part-period (EPP), planned capacity, and standard deviation of demands. The EPP directly affects the frequency of ordering and thus the lot-size. Twelve EPPs, 50 through 450, are applied to the lot-sizing techniques. Since a small value of EPP drives all lot-sizing techniques to LFL due to the low value of set-up cost, the lowest EPP considered was 50. Since 100 through 250 of EPP is the most sensitive ratio area, 12 EPPs(50, 100, 125, 150, 200, 225, 250, 300, 350, 400 and 450) are chosen for empirical investigation.

The demand data used in this chapter are the four types of requirements given in Table 10. These data were used in Berry [1] in a single-level system where no dependencies are given through lower level requirements from components required by the end item.

As previously stated, the components are assumed to be withdrawn from inventory at a uniform rate during each week. Therefore, the average inventory level will be used in computing the inventory costs. That is, inventory cost = (sum of beginning inventories + sum of ending inventories)(inventory carrying cost)/2. Set-up cost covers all fixed costs of ordering. It is calculated by multiplying the number of orders by the set-up cost for each item.

Week	Data sets			
	1	2	3	4
1	80	14	50	10
2	100	28	80	10
3	125	42	180	15
4	100	56	80	20
5	50	71	0	70
6	50	87	0	180
7	100	99	180	250
8	125	113	150	270
9	125	128	10	230
10	100	142	100	40
11	50	156	180	0
12	100	171	95	10
Mean	92.1	92.1	92.1	92.1
Standard deviation	27.0	51.4	66.1	130.1

Table 10: Demand patterns for investigation

Scheduling cost is directly related with the loading of equipment. It is assumed that there are two equipment types: an assembly-type and a machining-type equipment. The machine for assembly items has infinite capacity, but the machine for machining items has finite capacity. In other words, the machining-type equipment is restricted to a planned capacity which is a management specified plan for machine utilization. The planned capacity considered in this paper is either 40 hours or 50 hours for all 12 weeks. If the loads for processing exceed the planned capacity, production at an earlier date would be necessary. It will cause some inventory cost which will be a part of the scheduling cost. This cost will be called scheduling cost due to early production. Since the demand occurs at discrete periods and since all demands must be satisfied, demands which are needed for early production cannot be moved forward beyong the first week. That is, all demands should be satisfied, and therefore, for the first week it is assumed that planned capacity is unrestricted. An allowance for over-time work as part of scheduling cost is

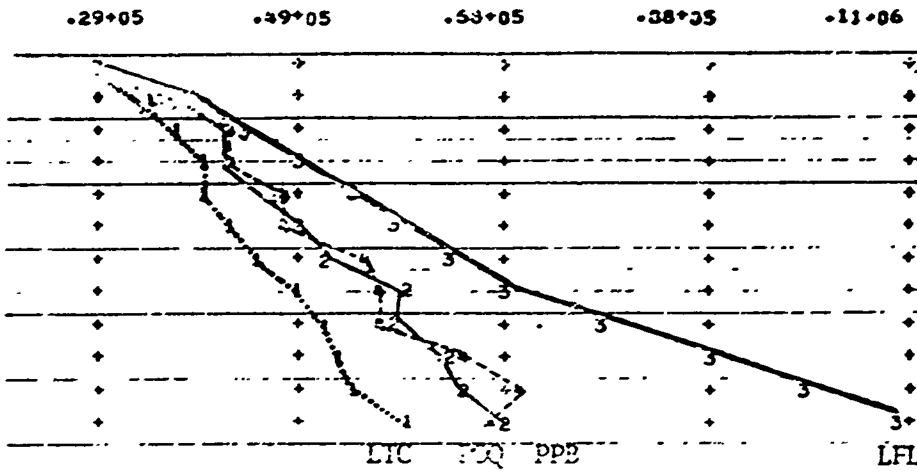
50% more than normal allowance, and over-time is the number of hours worked more than 40 hours per week. No penalty is imposed on idle time. It is also assumed that the labor wage per hour is $4. Thus, the total scheduling cost consists of the carrying cost due to early production and the allowance for over-time work.

3.1 Discussion of Computational Experiences

A computer program was developed to analyze the implications of the combined problem of lot-sizing and scheduling within a multi-level requirements planning system. The program was written to compute the total cost (inventory carrying cost, set-up cost and scheduling cost) of the five different techniques, and to determine which one is the most economical, for a given EPP ranging from a value of 50 to 450. Sixteen different runs were made to reflect varying manufacturing changes; demand changes (four demand patterns), product changes (two product diagrams) and planned capacities (two capacities, 40 & 50 hours). The resulting total cost plots, illustrated in Figures 2 & 3, are typical of the plots where eight runs are averaged for each product diagram.

For many of these runs, LTC resulted in the least total cost. The next higher cost procedure is POQ and PPB with no apparent difference existing between them. The next is LFL and the worst is the EOQ method. The EOQ method is restricted to fixed lot-sizes at non-uniform and discontinuous demands resulting in larger inventory costs. This occurs because of the mismatch between the order quantities and the demand values, which causes excess inventory to be carried forword from week to week and to be exploded from item to item of next level. For this reason, the total costs based on the use of the EOQ model are greater than the LTC, PPB, LFL and POQ models, and the results for the EOQ are not listed here, and will be excluded for further analysis.

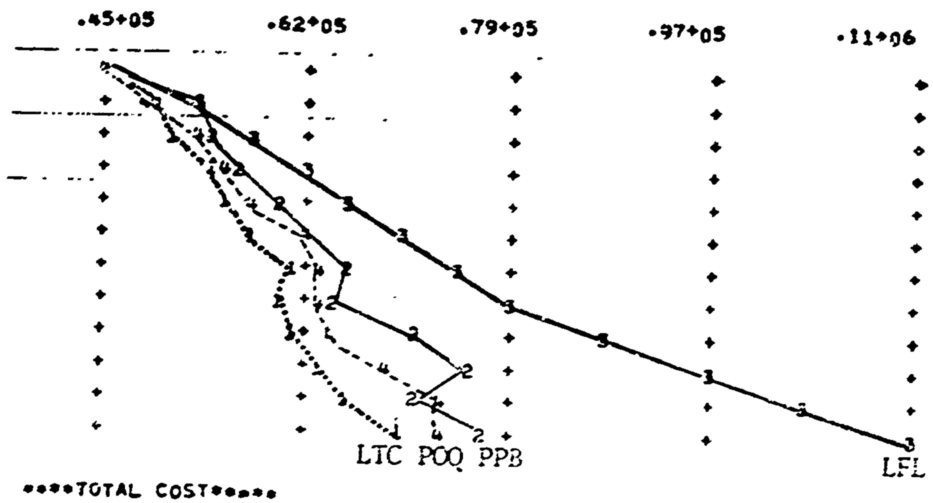
A comparison strictly on the basis of total cost may not be valid. A breakdown of the total costs into its component costs shows that set-up cost make up roughly 55% of the total cost with the standard deviation larger than the standard



MODEL 1 LTC	MODEL 2 PPA	MODEL 3 LFL	MODEL 4 PPB
[illegible]	.2553+03	[illegible]	.2955+05
[illegible]	.3243+05	[illegible]	[illegible]
.2634+05	.4236+05	.4410+05	.4238+05
[illegible]	.4302+05	.4093+05	.4262+05
.3983+05	.4537+05	.5033+05	.4790+05
.4243+05	.4920+05	.5363+05	.4352+05
.4503+05	.5226+05	.6350+05	.5531+05
.4925+05	.6000+05	.6835+05	.5671+05
.5127+05	.5872+05	.7305+05	.5727+05
.5251+05	.6377+05	.8775+05	.6431+05
.5473+05	.6465+05	.9745+05	.6971+05
.5960+05	.6815+05	.1072+06	.6771+05

****TOTAL COST*****

FIGURE 2. Typical Graph of Total Cost on Product Diagram



MODEL 1 LTC	MODEL 2 PPB	MODEL 3 LFL	MODEL [illegible] POQ
.4474+05	.4855+05	[illegible]	[illegible]
.4739+05	.5328+05	.5326+05	.6047+05
.5139+05	.5471+05	.5771+05	.5237+05
.5434+05	.5630+05	.6203+05	.5342+05
.5596+05	.6023+05	.6641+05	[illegible]
.5845+05	.6279+05	.7073+05	[illegible]
.6095+05	.6563+05	.7311+05	.6293+05
.6058+05	.6447+05	.7945+05	.6435+05
.6193+05	.7223+05	.8316+05	.6477+05
.5730+05	.7695+05	.9523+05	.6773+05
.5625+05	.7231+05	.1026+06	.7429+05
.7124+05	.7768+05	.1143+06	.7483+05

FIGURE 3 Typical Graph of Total Cost on Product Diagram 2

deviation for the other lot-sizing methods. This observation can be seen in Tables 11 and 12. It would appear then that the lot-sizing technique that reduces set-up cost may yield the best results. From this point of view, the LTC procedure is the best lot-sizing technique.

	Inventory cost	Set-up cost	Scheduling cost	Total cost
LTC	47.37%	47.16%	5.47%	100%($49751.09)
PPB	42.05	53.28	4.67	100 (55766.83)
LFL	33.80	62.67	3.53	100 (69645.02)
POQ	41.58	54.03	4.39	100 (56520.09)
Average	40.60	54.97	4.43	100 (57920.76)

Table 11: Percentages of each cost factor

	Inventory cost	Set-up cost	Scheduling cost	Total cost
LTC	$3856.28	$9777.05	$2361.64	$10072.38
PPB	4167.66	11938.64	2651.80	12243.72
LFL	4379.91	22548.20	2990.76	23859.93
POQ	3612.96	11212.52	2137.76	12462.31
Average	4004.20	13869.10	2535.54	14509.59

Table 12: Standard deviation of each cost factor

3.2 Statistical Inference

We should consider the results obtained from 16 runs with two product diagrams, two planned uniform capacities and four demand patterns, as statistical samples, and we can make some general conclusions based on these sample results through statistical inferences. The following statistical methods will be used for data comparison:

a) Multiple comparison technique

b) Regression analysis.

A multiple comparison technique called "Least Significant Difference(LSD)" is applied for comparing costs resulting from the different lot-

sizing procedures. Due to the fact that cost is a quantitative variable and the independent variables(EPP, standard deviation of demand, and planned capacity) are also quantitative, it would be possible to fit and analyze the response relationships between the costs and the independent variables through regression analysis.

Table 13 shows that the mean cost of LTC is significantly different from the PPB, POQ and LFL methods, because the difference between the mean of LTC and means of the other techniques is always greater than the value of LSD, 3079.10. Thus, LTC is the best lot-sizing method with a 5% Type I Error. LFL is shown to be the worst one among the four techniques. However, no apparent difference exists between the mean costs of POQ and PPB.

		LTC	PPB	POQ	LFL
Means		4951.09	55766.83	56520.09	69645.02
Differences	from LTC		6015.74	6769.00	19883.93
	from PPB			753.26	13878.19
	from POQ				13124.93

Table 13: LSD chart for comparison of total cost (LSD = 3079.10 with a 5% Type I error)

It is not sufficient to simply know that differences exist between the four lot-sizing techniques. It is equally important to know which of the component cost factors have significant impacts on the overall total cost. To obtain this information, regression analysis is applied to the data generated by the 16 runs. The independent variables considered are EPP, standard deviation of demands, planned capacity and product diagram. For EPP, standard deviation and planned capacity, the actual values are used, while 0-1 coding is adopted for the two product types. The SPSS package [6] is used for getting the results of the regression analysis.

First, we examine whether the multiple linear regression model is adequate to explain the functional relationship between the total cost (Y) and the independent variables, X_1 to X_4. The fitted model is:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

where X_1 = EPP-ratio

X_2 = standard deviation of demand

X_3 = planned capacity

X_4 = product diagram

β_i = regression coefficient of i^{th} independent variable

ε = error term.

The regression equations with R^2 (coefficient of determination) obtained from the SPSS packages are shown in Table 14. The high R^2 values reveal that the multiple linear regression equations are adequate to explain the total cost through the independent variables.

	Regression equations
LTC	$Y=3422.85+1784.69X_1+75.48X_2+63.30X_3-178.24X_4$
PPB	$Y=37151.20-111.57X_1+94.21X_2+74.84X_3-209.45X_4$
LFL	$Y=26324.84+485.58X_1+188.77X_2+93.70X_3-190.91X_4$
POQ	$Y=33400.73+1778.33X_1+90.90X_2+77.41X_3-124.33X_4$

	R^2
LTC	.867
PPB	.991
LFL	.958
POQ	.829

Table 14: Regression equations and R^2

To determine which factors are important (or not important) to each of the four lot-sizing techniques, we can test the null hypothesis $H_o : \beta_i = 0$. The test statistic is $t = b_i/s_{b_i}$, where b_i is the estimated coefficient of regression of β_i and s_{b_i} is the standard deviation of b_i. The t values are listed in Table 15.

Since the critical value of t is 1.072 with 5% level of significance, we observe that X_1 and X_2 are the most important variables for all techniques as shown in Table 16. This means that one should pay close attention to these two variables, no matter what lot-sizing technique is used. Among these two variables, EPP-ratio(X_1) turned out to be more important than the standard deviation of demand(X_2).

Lot-sizing technique	EPP(X_1)	Standard deviation(X_2)	Planned capacity(X_3)	Product diagram(X_4)
LTC	34.97	10.88	3.51	3.53
PPB	41.82	12.33	3.93	.21
LFL	64.21	11.79	2.74	.70
POQ	28.57	8.99	1.65	2.36

Table 15: T test values

Lot-sizing technique	Important variables in order
LTC	X_1 X_2 X_4 X_3
PPB	X_1 X_2 X_3
LFL	X_1 X_2 X_3
POQ	X_1 X_2 X_4

Table 16: Important variables in order for each technique

It is further observed that the product diagram (X_4) does not significantly affect the total cost

for the PPB and LFL techniques. As for the POQ method, planned capacity is not a significant factor.

We have analyzed the relationship between the total cost and four factors; EPP-ratio(X_1), standard deviation of demand(X_2), planned capacity (X_3) and product diagram(X_4). A similar analysis can be done for the inventory cost, the set-up cost and the scheduling cost, respectively, with respect to these four factors.

Frequently, the inventory carrying cost is the major cost factor in the operation of a plant. For such cases, it might be desirable to compare lot-sizing techniques only by inventory carrying cost. No conclusion could be made on the best lot-sizing technique for inventory carrying cost using the LSD test. However, based on the results of the regression analysis, the standard deviation of demands significantly affects the cost of carrying inventory. Therefore, plants whose main concern is inventory carrying cost should pay close attention to the demand pattern. For details, see Park [5].

As discussed before, the set-up cost might dominate the total inventory cost. It is also valuable to compare lot-sizing techniques only by set-up cost. The LSD test for set-up cost indicates that LTC is the best method, and PPB is the second best technique. The next best one is POQ, and LFL is the worst one among the four techniques. By regression analysis, EPP-ratio(X_1) is the most important factor for set-up cost, and then standard deviation of demands(X_2). The other factor, X_4, is also an important factor to LTC, LFL and POQ.

Set-up cost and inventory carrying cost have a strong relationship with each other in terms of the variables X_1,X_2 and X_4. But variations in wage rates, over-time allowance and increases or decreases in planned capacity may not affect the set-up cost and inventory cost. So, it may also be important to compare the four techniques by the scheduling cost. Using the LSD test for scheduling cost, no significant difference is observed between LFL and POQ, but these two methods yielded the lowest scheduling cost. The second best is the PPB, and LTC is the worst one. The EPP-ratio does not affect the scheduling cost

by the regression analysis, but this cost is affected by changing the standard deviations of demands. X_3(planned capacity) and X_4(product diagram) are also important factors.

REFERENCES

[1] Berry, W.L., LOT SIZING PROCEDURES FOR REQUIREMENTS PLANNING SYSTEMS: A FRAMEWORK FOR ANALYSIS, Production and Inventory Management, Vol. 13, No. 2., pp. 19-34, 1972

[2] Gorham, T., DYNAMIC ORDER QUANTITIES, Production and Inventory Management, Vol. 9, No. 1, pp. 75-81, 1968

[3] IBM Data Processing Manual, THE PRODUCTION INFORMATION AND CONTROL SYSTEM, pp. 46-58, 1968

[4] Orlicky, J., MATERIAL REQUIREMENTS PLANNING, MaGraw-Hill Book, Inc., New York, 1974

[5] Park, E.H., LOT-SIZING AND MACHINE LOADING IN MULTI-LEVEL REQUIREMENTS PLANNING SYSTEMS, Unpublished Master Thesis, Mississippi State University, 1977

[6] STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES (SPSS), McGraw-Hill Book Company, New York, 1975

[7] Theisen, E.C., NEW GAME IN TOWN, THE MRP LOT SIZE, Production and Inventory Management, Vol. 15, No. 2, pp. 1-13, 1974

[8] Wagner, H.M. and Whitin, T.M., DYNAMIC VERSION OF THE ECONOMIC LOT SIZE MODEL, Management Science, Vol. 5, No. 1, pp. 89-96, 1959

A GENERAL PURPOSE SIMULATION SYSTEM AND ITS APPLICATION WITHIN A MANUFACTURING COMPANY

G.D. Craig

Department of Management Studies

University of Waikato

Hamilton

New Zealand

ABSTRACT

The purpose of this paper is to describe the structure and use of a general purpose simulation package developed by the author, for the interactive construction of dynamic computer-based simulation models. The package has evolved from several years experience gained in developing corporate planning procedures in a variety of companies.

The package consists of a system of computer programs written in the 'BASIC' language, for a Digital PDP 11/70 time-sharing computer. The framework upon which the package is based, is essentially an amalgamation of the System Dynamics and Input-Output approaches to the modelling of complex organisations. Networks of levels and flows provide the basis for dynamic representation, while matrix algebra features provide the basis for aggregational flexibility, and the analysis of multi-product, multi-process production systems.

The structure, user procedures, capabilities, and limitations of the package generally are discussed, together with some practical aspects of its use within a manufacturing company as a corporate planning tool. A methodology for setting up formal corporate planning procedures within a small company is also discussed, together with the structure and use of three specific planning models developed by the author for the company.

1. INTRODUCTION

1.1. Purpose of the System

GENSIM, (for Generalised Simulation System), is a computer package, or system of programs, designed to support the interactive construction and use of dynamic simulation models. It consists of a suite of seven programs, (refer Figure 1), written in the 'BASIC' language for a Digital PDP 11/70 computer system under the RSTSE operating software.

GENSIM has been designed primarily for general purpose modelling in a corporate planning context, i.e. the construction and use of flexible dynamic models encompassing those aspects of the corporate entity which can be expressed in quantitative terms; either in isolation, or as an integrated whole. Thus, for example, financial models, marketing models, production models, and purchasing models can be constructed, either as separate 'stand-alone' models to support the planning of specific subsystems of the corporate system, or as one integrated 'corporate' model supporting the quantitative side of corporate planning generally.

In both instances the package is designed to provide management with a facility for quickly testing the broadest possible range of strategies, from those involving merely scalar or parametric changes in a system of given structure, to those involving fundamental changes in the system structure, or even entirely new systems such as new business ventures.[1]

---

[1] See Naylor and Mansfield (4) for suggested criteria for the design of computer-based modelling systems. The author does not agree however, with the argument advanced therein for simultaneous equation solution capability. Dynamic, recursive modelling systems utilising feedback loops can adequately cope with the kind of 'simultaneity' described in the examples cited on p.20. The five equation financial model described by Naylor and Mansfield is a single period model in which simultaneity arises from the assumptions upon which the model is based, rather than from the real world system being represented. Specifically the model assumes that both tax and interest occur as cash flows, in the same period as the amounts are charged against income. Extension of this model to permit dynamic representation. with a suitable choice of solution interval not only removes the simultaneity, but also improves the model's validity.

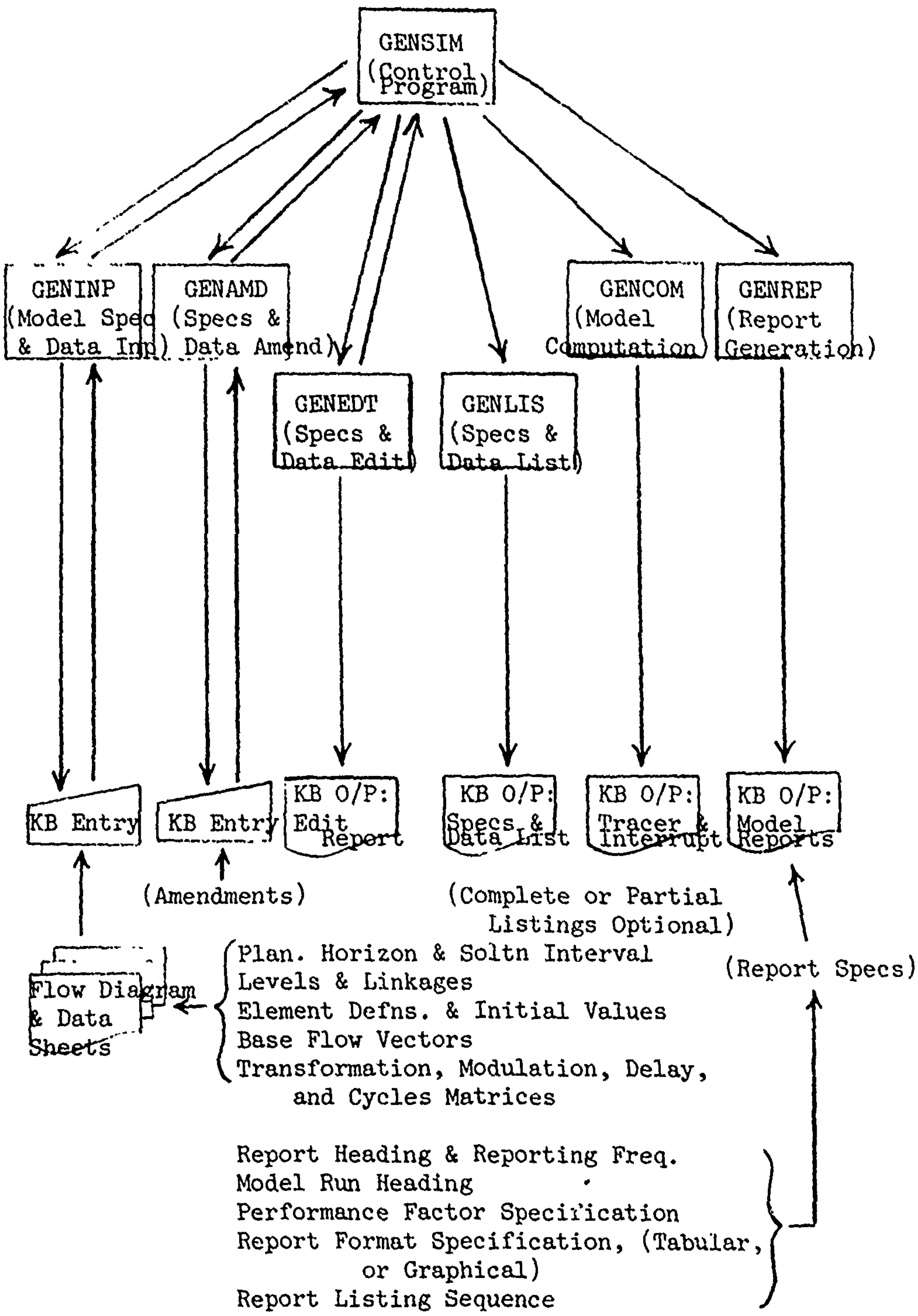


Figure 1. GENSIM System Programs

1.2. Corporate Planning and the Concept of a Goal Structure

Figure 2 sets out in general terms, the nature of the corporate goal structure, and its dependency upon change achieved through the implementation of co-ordinated actions, which in turn should evolve from consideration of the goal structure. Ideally, the corporate planning process should result in the specification of a congruent goal structure, supported by a set of integrated action plans, which are then monitored in terms of the actual changes which they induce when implemented.[2] The whole process is of course immensely complicated by the presence of uncertainty and a multitude of external influences which determine boundary conditions with respect to goals and actions, affect human behaviour, and induce 'unplanned' changes in the system and its environment.

1.3. Corporate Planning in Practice

The above discussion leads to the identification of four major areas of difficulty associated with corporate planning in practice.

1. How to specify a congruent goal structure.
2. How to develop integrated, and adequate action plans to support the goal structure.
3. How to maintain the relevance and validity of the goal structure and action plans in the face of the 'unplanned' changes.
4. How to gain managerial acceptance of, and commitment to, the corporate plan.

It is submitted that the most important single factor contributing to the first three of the above (and probably indirectly pertinent to the fourth also), is often the lack of ability on the part of the planners to efficiently perform computations involving a large number of factors, and complex interdependencies. This is reflected in the fact that for even the most

[2] A cyclical process within which four distinct phases can be identified, along the lines suggested by Perrin (5) Ch.2.

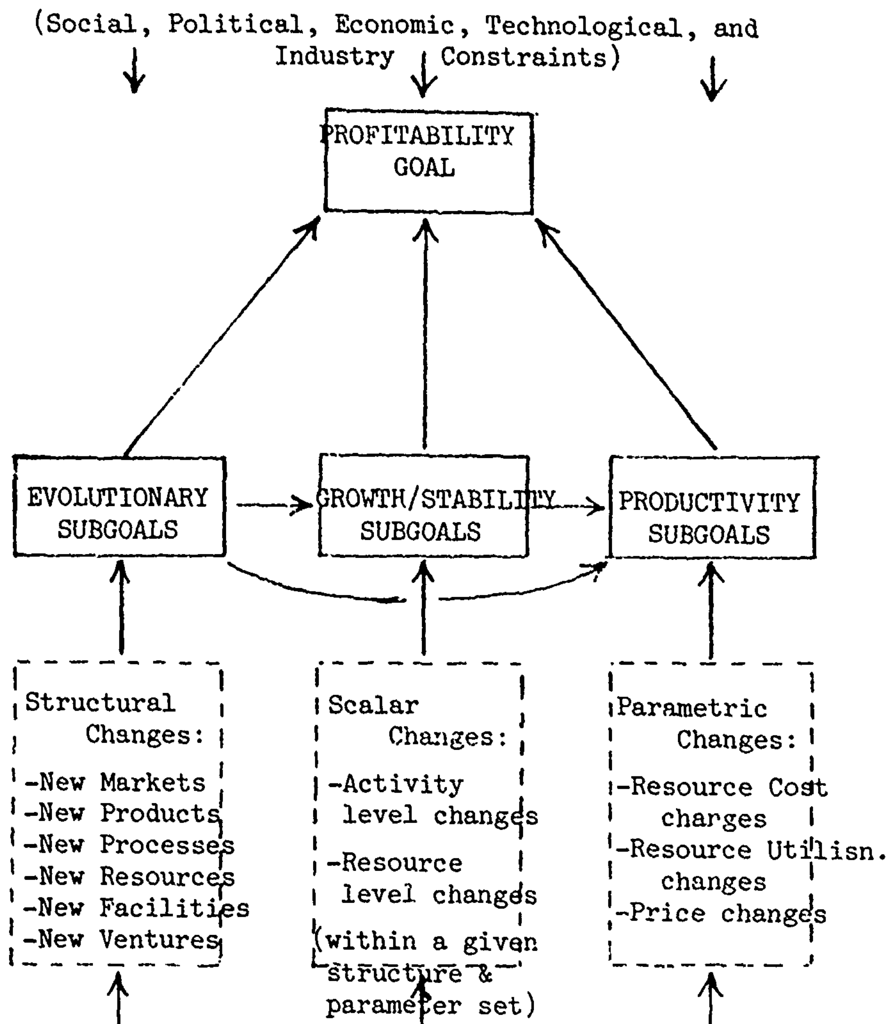


Figure 2. The Integrated Corporate Goal Structure

advanced organisations, the corporate planning process has not developed much beyond 'single-future' planning strait-jacketed into a 'fixed interval-periodic review' format, (usually 5-year planning with annual reviews). Even current literature offers little insight into the nature of the next stage in the development of corporate planning from its humble 'ad-hoc' beginnings, through the twelve month budgeting phase, to the present day.[3]

The above inadequacy drastically limits the strategy search and evaluation phases, the extent to which plans can be adapted to reflect 'unplanned' changes, and hence the overall utility of the plans and their acceptability in the eyes of managers. It does little to foster the creative thinking on the part of managers which constitutes the source of all impetus for planned change. As an essentially spontaneous process, creative thinking should not be confined merely to annual 'brainstorming' sessions.

The GENSIM package is designed to provide, as far as possible a tool for overcoming this deficiency, and for facilitating -

1. the formulation of congruent goal structures and co-ordinated action plans, through consideration of the widest possible range of strategies within and explicitly defined 'systems' framework, which can be readily flexed in response to new developments (both internal and external to the organisation).

2. The shift from 'fixed interval-periodic review' planning of a single future, to 'flexible interval-continuous review' planning of multiple futures.

2. UNDERLYING CONCEPTS AND STRUCTURE OF GENSIM

2.1. System Dynamic and Input Output Analysis

The general framework upon which the package is based is essentially an amalgamation of elements from two 'systems approaches' to the modelling of complex organisations; that of System Dynamics, and that of Input-output Analysis .

---

[3] Cantley (2) addresses this question, but confines his prognosis to types of models, rather than the process itself, and the role of models in enhancing it.

The former represents organisations in terms of information feedback loops embodied in a web of inter-connected networks of levels and flows pertaining to orders, materials, money, equipment, personnel, and information. The management process is seen as being a process of influencing levels by regulating flow rates on the basis of information feedback, over a continuum of time intervals.[4] An important consideration with models constructed on this basis, is the adoption of a level of aggregation of model variables sufficiently high to permit valid representation in a continuous flow model capable of depicting organisational growth and stability characteristics.

From the corporate planning viewpoint, the System Dynamics approach appeals mainly because of the extent to which it permits dynamic representation of an organisation, as a 'total system'. Its principal disadvantage perhaps, is that the level of aggregation of System Dynamics models is usually such that their usefulness is restricted to the growth and stability aspects of the organisation. Their utility for comprehensive corporate planning encompassing the kind of goal structure depicted in Figure 2, is limited. The development and evaluation at least in profitability terms, of strategies involving evolutionary and productivity facets of the corporate goal structure, and also the need to translate strategic plans into tactical plans, requires the support of a less aggregated, and more flexible, modelling approach.

Input-output analysis utilises the power of matrix algebra to represent the quantitative aspects of organisations in terms of sectors, with complex inter-sector interdependencies being accommodated through the use of matrix transformations.[5] Differing levels of aggregation can be adopted, within the constraints

---

[4] See Forrester (3), p. 95.

[5] See Butterworth and Sigloch (1) for an excellent rationalisation of a range of examples of Input-output Analysis at the micro-economic level, in terms of a generalised Input-output model.

imposed by the assumptions of linearity and coefficient stability, but the main advantage of the technique, from the corporate planning standpoint, is its flexibility and power in reflecting complex multi-product, multi-process interdependencies.

The over-riding disadvantage of traditional Input-Output models, however, is that they constitute a static form of analysis, reflecting flows only, over a single period.

2.2. The Elements and Framework of GENSIM

The systems framework upon which the GENSIM package is based incorporates elements of both the System Dynamics, and Input-Output frameworks. The networks of levels interconnected by flows of the former provides for dynamic representation , while the matrix algebra of the latter provides for aggregational flexibility, and the ability to set up and use more disaggregated models efficiently. This hybrid framework may be summarised in the following terms:

1. The fundamental structural elements of the framework are level vectors and flow vectors.

2. The level vectors are related, where appropriate, by linkages, with each linkage consisting of an inflow vector, outflow vector, and transformation matrix - for transforming the inflow into the outflow or vice versa. Each linkage is classified according to whether the transformation is general, sum, or identity.[6]

---

[6] This in effect is the application of Input-Output Analysis at the 'ultra-micro' level, with the simultaneous equation framework defined within the linkage and solution interval. Both primal and dual forms of the transformation can be evaluated for any linkage.

3. Dynamic representation is achieved through the use of a suitably defined model solution interval,[7] and the balancing relationship -

(Level vector values at interval end) = (Level vector values at interval start)

+

(Sum of inflows)

-

(Sum of outflows)

4. System interdependencies are reflected through the specification of a <u>flow dependency type</u> classification for every flow vector element in the model, chosen from a range of options whereby a <u>flow determinant</u> is selected. Possible flow determinants are level vector element, inflow vector element, outflow vector element, or exogenous time-varying influence.

5. Time triggered influences such as delays are reflected through the specification of a <u>flow pattern classification</u>, for every flow vector in the model, from a choice of four basic options which govern response of the dependent flows to their determinant factor, over time. These options are instantaneious response, exponential delay, distributed box car delay,[8] and stackup boxcar delay.[9]

---

[7] Unlike System Dynamics models, <u>flows</u> for each solution interval are calculated, rather than <u>flow rates</u> - so that non-linear behaviour is represented by a series of broad linear approximations. Obviously the finer points of a system's dynamic characteristics are masked by this approach, but the ability to handle more disaggregated models which this facilitates, (by reducing the computational load to one iteration, or solution interval, per period), more than compensates for this as far as the utility of the system for corporate planning is concerned.

[8] Where a flow is distributed across a series of future periods.

[9] Where a flow is accumulated for a specified number of periods then the accumulated flow is distributed across a series of future periods.

6. Exogenous influences are reflected through the use of <u>modulation matrices</u>, and <u>cycle vectors</u>. In the case of the former up to two matrices may be tagged against any flow vector so that each element of the flow vector is paired up with the corresponding row of the modulation matrix, which carries the code for any one of 10 possible flow modulating patterns, (steps, pulses, ramps, etc.). This number of patterns is effectively doubled because any pattern can be coupled up with a 'seasonal index' series chosen from among the cycle vectors, (each vector carries the indices and cycle periodicity of a series).

2.3. Distinctive Features of GENSIM

The general systems framework described above provides some distinctive features which are not found in conventional 'corporate' modelling systems, many of which are based on the less structured and more restrictive accounting framework. These features are summarised below:

1. Both financial and non-financial aspects of an organisation can be modelled with equal ease - either separately or as an integrated 'corporate' model.

2. The System Dynamics concepts permit the representation of complex time-lagged relationships and flexibility in the choice of solution interval, to avoid masking the dynamic characteristics of a system through conformi to an 'accounting period' solution interval such as one year.

3. The Input-Output concepts incorporating the use of matrix algebra greatly facilitate the representation of complex multi-product multi-stage production systems in both a primal and dual form. The primal form accomplishes the stage-wise explosion of final demand into resource input demand for each solution interval, while the dual form (automatically derived from the primal form), develops the unit cost build-ups for the final products based on a given set of resource input unit costs, for these aspects of the system for which the concept of unit cost applies.

4. Models are constructed, edited, computed, and reported on in a fully interactive manner, without the need for any programming or entry of lines of code. All of the information necessary for specifying model relationships and model reports, together with the basic data for the model computation, is entered at the computer terminal by the user, in response to clearly worded prompts which specify the options available to the user, where appropriate. The model development phases which the user must complete unaided are confined to:

(i) Development of the flow diagram specifying the system levels and flows as an interconnected network, together with the location of the required matrix transformations, and the flows to be initialised, modulated, and delayed.

(ii) Determination of the necessary data to specify each transformation matrix, flow modulation pattern, flow delay pattern, and flow initial values.

(iii) Determination of the necessary data to define the specific measures of performance which the user wishes to have reported, and the manner in which they are to be reported.

Excerpts from the computer terminal conversation associated with the setting up of the first model discussed in this paper, are set out in Appendix I.

3. USING THE SYSTEM

3.1. The System Programs

The seven programs constituting the GENSIM package are shown in terms of their functional roles and interrelationship, in Figure 1. The most important feature of the system for the user is that models are developed and used interactively at the terminal, and that no programming in the conventional sense, is required to develop a model. Each program in the system performs its function according to a defined sequence of prompts and user responses. The command 'RUN GENSIM' causes the control program to be executed, allowing the user to choose any one of the six functional programs comprising the system. Each of these programs are discussed below in the sequence in which they would normally be accessed.

3.2. Setting up a Model

The setting up of a model is carried out using the 'GENINP' program, in two distinct phases. These phases are described below in their normal order of execution, although each can be performed independently if desired, for the purpose of effecting changes in a model subsequent to its initial specification.

1. Input of Model Specifications - where the model structure is specified, in response to the appropriate prompts requesting:

    the number of levels
    the number of linkages
    the type of model solution interval
      (days,weeks,months,quarters,years)
    the duration of the planning horizon
    the characteristics of each level
    the characteristics of each linkage

2. Input of Model Data - where the data relevant to the specified model is entered in response to the appropriate prompts. The user may select from among seven distinct entry points, or proceed through them sequentially. In each instance the levels (or linkages), are scanned and prompts requesting the appropriate data are generated accordingly. The entry points are for:

    level initial values
    flow initial values [10]
    transformation matrices
    modulation matrices
    delay matrices
    cycles

   In the case of the transformation and modulation matrices, data entry is facilitated by the ability to automatically zero fill or copy any row from a previously specified matrix. The 'GENAMD' program may be used for small scale amendments to the model structure or data, although specific levels, linkages, and matrices can be accessed via 'GENINP' for this purpose. Both programs chain the user back to the control program 'GENSIM' after execution.

---

[10] Base flow vector values must be specified for all exogenously dependent flows.

3.3. Editing the Model Structure and Data

Following model specification and data input, the 'GENEDT' program would normally be accessed to check out the logical consistency of the model structure, and the validity of the data submitted on the basis of this structure. This program scans each linkage for discrepancies or inconsistencies in the responses entered during the model specification phase, and prints out appropriate error messages. If desired the program will also scan all of the model data arrays for discrepancies or the prescence of values outside user-specified upper and lower bounds.

A visual check of all specifications and data is usually advisable prior to running the model, and to this end the data listing program 'GENLIS' would be accessed. This program takes the user through a series of steps to specify a data listing, as full or partial listings are available at the option of the user. A data listing file is created which the user can store and access as desired.

3.4. Model Computation

Model computation is performed using the 'GENCOM' program which allows the user to re-specify the planning horizon, to monitor the computations via a 'tracer', or to interrupt the computations at pre-defined points. The tracer provides on-line feedback, through the terminal, of each computational step in terms of equation type, calculated values, and supporting data pertaining to the values of relevant tag numbers, indicators, and counters. If desired this information can be restricted to one nominated linkage. In general however this feature constitutes an invaluable desk checking aid, as well as a supplement to the main system reports. After processing the model 'GENCOM' provides the user with the option of either chaining back to the control program, or proceeding directly to the report generating program 'GENREP'.

3.5. Model Report Generation

The report generating program 'GENREP' takes the user through a sequence of steps which constitute four phases described as follows:

1. Report Identification - where the model name and run heading is specified.

2. Report Performance Factor Specification - where the user can develop formulae to calculate up to 40 measures of model performance, based on level vector element values combined as ratios, percentages, or dimensioned quantities such as the sum of several level elements.

3. Report Format Specification - where the user can select from four basic options, as follows:

   tabular with report columns designated as vector elements

   tabular with report columns designated as time periods

   graphical with up to 10 graphs each plotted with up to five factors

   a combination of both tabular and graphical reports

4. Report Content - where the user specifies, in the desired sequence, the level vectors and performance factors which are to be reported together with the location of spacing and report headings.

Upon completion of the above phases, 'GENREP' then generates the desired reports using the file of computed data produced by 'GENCOM', and creates a report file for storage, or immediate access by the user.

4. APPLICATION OF THE SYSTEM IN A MANUFACTURING COMPANY

4.1. The Company

Hallmark Industries Ltd. is a relatively small privately owned company located in Hamilton, New Zealand. It specialises in the manufacturing of a compact range of tents, back packs, sleeping bags, and camping accessories , in what is for the most part a two-stage production process. The company employs a staff of approximately sixty-five people with the ownership of the company being vested in seven shareholders, five of whom hold key positions in the management structure. Approximately 60% of the company's raw material requirements are imported, along with a range of supplementary products which the company markets under license, as a small but significant part of its operations.

Over the past seven years the growth of the company has been extremely rapid, to the extent that both sales and total assets have increased six-fold. Profitability in terms of return on equity, although high, has declined somewhat due to a sixteen-fold increase in the company's equity base arising from the retention of earnings over this period. This high rate of growth, coupled with a marked seasonality in sales, has resulted in recurring liquidity and stock control problems together with undue reliance on costly short-term finance and difficulties in adapting the organisational structure and management style of the company to meet the changing scale and nature of its operations.

In recognition of these problems, formalised corporate planning procedures were instituted two years ago, based upon the traditional framework of five-year planning with annual reviews, and focussing on a goal structure defined in terms of the key factors set out in Figure 3. The annual reviews were carried out by the executive shareholders, in accordance with the following sequence of steps.

1. A broad review of the previous year's corporate plan in the light of actual performance against it.
2. The tentative re-specifying of the five-year company goals.
3. An Environmental Audit, to establish threats and opportunities in the social, legal, political, economic, technological, and industry environments.
4. A Company Audit, to establish the strengths and weaknesses in the functional areas of the company administered by each of the executive shareholders.
5. The development of functional area plans, comprising functional area goals and key actions for the attainment of the five-year company goals
6. Final specification of a vertically and horizontally integrated goal structure spanning the five years, with the goals for the immediate year ahead being expanded into monthly budgets and action plans, to provide the basis for control.

The introduction of a flexible, dynamic modelling facility in support of this process has resulted in the quantitative assessment of a much broader range of strategies in the course of conducting steps 5 and 6 above, together with a

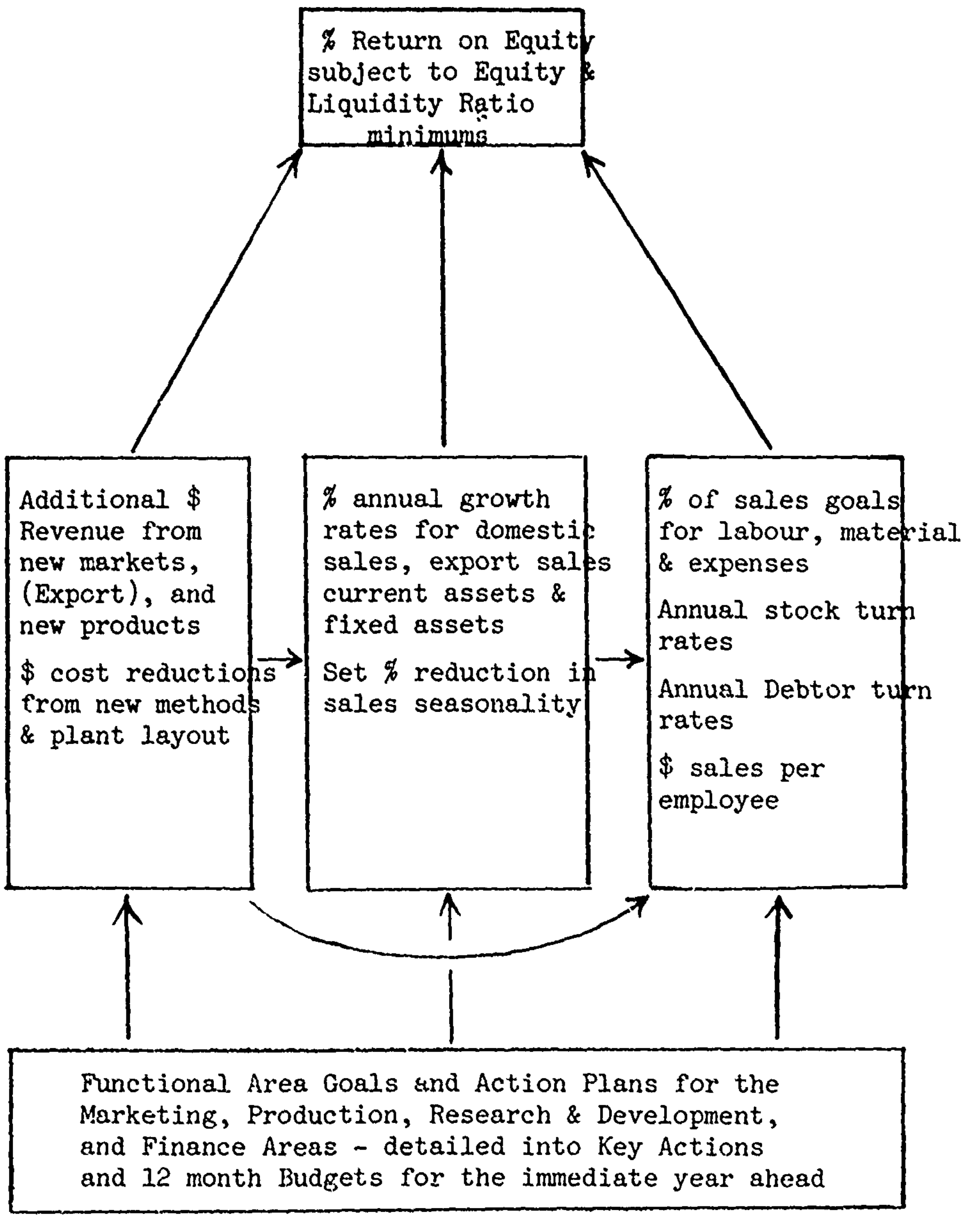


Figure 3. Corporate Goal Structure for Hallmark Industries Ltd.

move away from 'single future-annual review' planning to the type of 'multiple future-continuous review' planning referred to in Section 1.3. At present three models provide the basis for this development, viz., a Cash Planning Model, a Production Planning Model, and a Corporate Planning Model. The structure and purpose of each of these models is discussed in the remainder of this section.

4.2. The Cash Planning Model

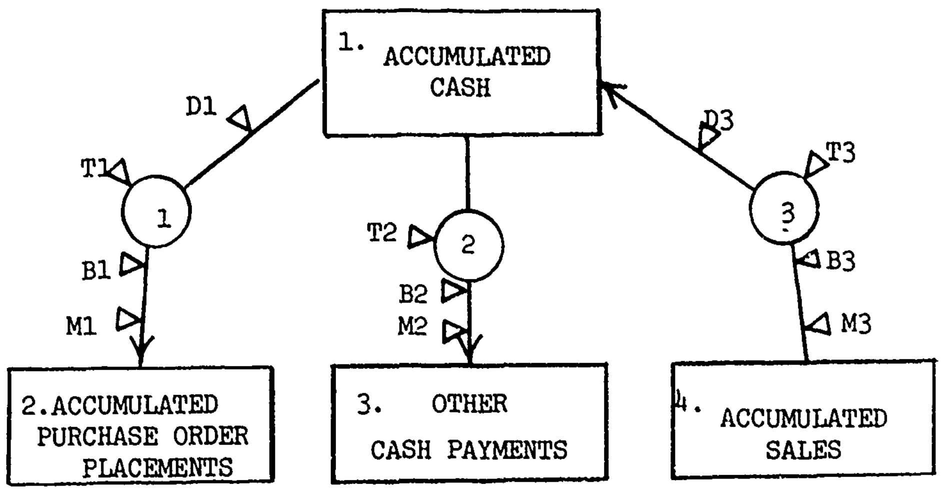
This model was the first to be developed as the underlying relationships associated with cash flows within the company are typically more structured, and the benefits of such a model are more immediate and visible to management. Figure 4 depicts the basic structure of the model as a flow diagram, and the computer terminal conversation associated with its development using GENSIM is set out in Appendix I.

The model encompasses both cash flows, and those flows which have a direct bearing on them such as sales, and the placement of orders for materials. Two delays are incorporated into the structure to reflect the significant lag effects associated with cash collections from debtors, and the payment of bills associated with imported materials purchases. The model is structured around a solution interval of one month with the facility to project up to sixty months ahead. The model can be initialised at any given month of the year to provide, for example, rolling twelve month projections.

4.2.1. Model Structure

Figure 4 depicts the model using a set of symbols which can be adopted for any model to be constructed using GENSIM. All levels defined in the system are represented by rectangles, and all flow linkages by arrows with the linkage number noted in a small circle which splits the linkage into an inflow and an outflow respectively.It is important to note that the linkages are multi-channel with each outflow (or inflow) possessing the number of channels equal to the dimension, (i.e. number of elements), defined for the level which it leaves (or enters).

The existence of external influences on the flow channels and the switching, or transformation matrices relating one side of a linkage to the other, are indicated by the small triangles, or tags. These tags are identified by one of four letters, (B, M, T, or D), followed by a number. The letters denote respectively base flow vectors, modulation



Level Element Definitions

Accumulated Cash	El.1. Local Collections 2. Export Collections 3. Local Purchases 4. Import Purchases 5. Other Payments
Accum. Purch. Orders	El.1. Local Orders Placed 2. Import Orders Placed
Accum.Other Payments	El.1. Manufacturing Wages 2. Overheads 3. Salaries 4. Capital Expenditure 5. Repayments/Borrowings 6. Tax Payments 7. Distributions
Accumulated Sales	El.1. Local Sales 2. Export Sales

Figure 4. The Cash Planning Model

matrices, transformation matrices, and delay matrices, while the associated number constitutes a tag number which enables identification of the particular data array which is to be accessed to give effect to the initialisation, modulation, transformation or delay.

In this particular model three base flow vectors, three modulation matrices, three transformation matrices, and two delays have been defined. The data arrays associated with these influences constitute part of the data input to the model.

4.2.2 Model Input

The following inputs must be supplied to the model. It is important to note that any data item so supplied, may be treated as a constant, or as a decision variable, (i.e. parameter), depending upon the circumstances governing the use of the model.

1. Level Initial Values - zero in all instances for this model as the levels represent accumulated flows.
2. Flow Transformation Matrices -

   T1 switches the purchase orders placed flows into the purchases payments flow channels

   T2 switches the other cash payments classifications into the other cash payments flow channel.

   T3 switches the sales flows into the cash collections flow channels
3. Base Flow Vectors - which comprise the base, or initial month values for the flows in each one of the flow channels of the inflows (or outflow) to which they are assigned.
4. Modulation Matrices - which act on the flows in each channel of each of the inflows (or outflow) to which they are assigned, to increment them in accordance with individually specified growth patterns.
4. Delay Matrices -

   D1 constitutes a boxcar delay reflecting the combined lag effects of materials delivery lead times, and payment terms, (120 day bills in the case of imports).

   D2 constitutes a boxcar delay reflecting the lag in cash collections from debtors.

5. Cycles Matrix - which comprises ten uniquely defined cyclical series, each of which is coupled up to a specific flow channel's modulation pattern. The flows being subjected to cyclical or seasonal influences in this way include sales, purchase order placements, tax payments, and manufacturing wages.

4.2.3. Model Output

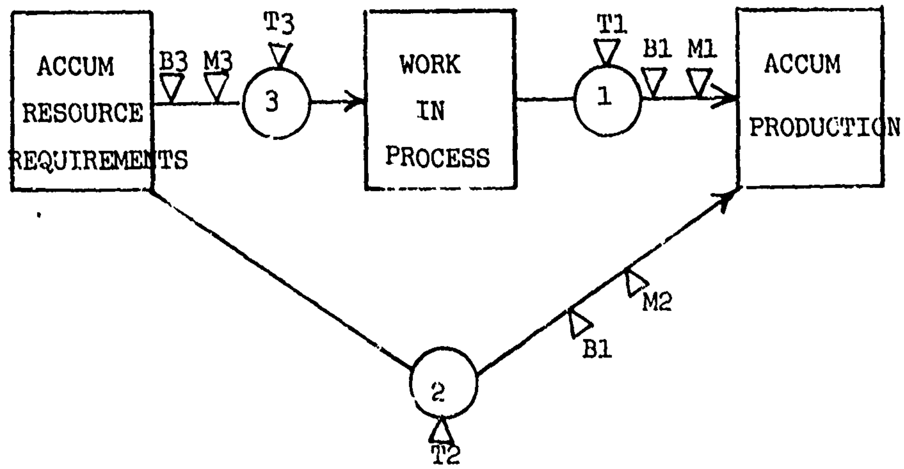
Sample output, in tabular form, from one run of this model is shown in Appendix II. The particular run shown here is a twelve month projection commencing with the month of October. In practice management run this model at least twice per month, (to produce optomistic and pessimistic projections respectively), with special runs over and above these to determine the cash effects of revised capital expenditure programs, or unforeseen events such as devaluation for example.

4.3. The Production Planning Model

This model utilises the full power of the matrix algebra features of GENSIM in performing the transformation of final product demand, (i.e. production levels), into resource input requirements, together with the dual transformation of resource input unit costs into final product unit costs. A solution interval of one year is used with a planning horizon of five years, to interface with the Corporate Planning Model.

4.3.1. Model Structure

The basic structure of the model is depicted in Figure 5 using the symbols described earlier for the Cash Planning Model. Two of the three levels of the model represent accumulated flows (for production and resource requirements respectively), while the Work-in-Process level exists basically as a dummy level to permit the separate representation of the flows associated with each stage of the two stage production process by which some of the products are manufactured. Unlike the Cash Planning Model, (and the Corporate Planning Model), these levels are defined in physical terms, with conversion to monetary units being effected via the associated unit cost vectors. The primal form of the model performs the final product - resource input transformation, for each year of the planning horizon, while the dual form effects the resource unit cost - final product unit cost transformation.



Level Element Definitions

Accumulated Resource Requirements El.1. Pack Nylon
2. Tent Nylon
3. Sleeping Bag Fill
4. Pack Webbing
5. Metal Tubing
6. Miscellaneous
7. Metalwork Man-hrs
8. Packs Man-hrs
9. Tents & Other Man-hrs

Work-in-Process El.1. P1 Pack Frames
2. P2 Pack Frames
3. Pack Nylon
4. Pack Webbing
5. Miscellaneous
6. Packs Man-hrs

Accumulated Production El.1. P1 Packs
2. P2 Packs
3. T1 Tents
4. T2 Tents
5. T3 Tents
6. SB1 Sleeping Bags
7. SB2 Sleeping Bags
8. Accessories

Figure 5. The Production Planning Model

Three transformation matrices carry the resource consumption coefficients associated with the manufacturing processes, while the base flow vectors initialise the production flows (of final products), and the resource unit costs, for the first year. The modulation matrices then act upon these initial values to give effect to the desired growth patterns for final production and resource unit costs.

4.3.2. Model Input

The inputs required for this model are summarised below. As with the Cash Planning Model, any individual data item (or items), may be selected as a parameter.

1. Level Initial Values - zero in all instances as levels 1 and 3 are accumulated flows, and level 2 is a dummy level.
2. Flow Transformation Matrices - T1 carries the resource consumption coefficients associated with the production of those products made in a single stage, (linkage 1), T2 carries the coefficients relating to the second stage of production, (linkage 2) for those products utilising two stages. T3 carries the coefficients for the first stage, (linkage 3).
3. Base Flow Vectors - B1 is the production base flow vector, and B3 is the resource base unit cost vector.
4. Modulation Matrices - M1 carries the production growth patterns for each final product, over the five year planning horizon, and M3 carries the 'growth' pattern for the unit cost of each resource input. These patterns are in fact annual percentage increases in each case.

4.3.3 Model Output

Sample output, in tabular form, from one five year run of this model, is shown in Appendix III. In practice this model is used by management for both the short, and long term planning of production. In the former instance, a version of the model using a monthly solution interval is run each month on a rolling twelve month basis, using input from the Corporate Planning Model, to provide detailed resource input requirements as a basis for a detailed purchase order placements plan. The dual form of unit cost buildups also provides a basis for short-run action plans aimed at production cost reduction.

For the longer term, the yearly solution interval version serves as a susidiary model to the more comprehensive Corporate Planning Model discussed below.

4.4 The Corporate Planning Model

This model plays a central role in the planning process and is of necessity the most comprehensive of the three models. In its present form it encompasses all financial aspects of the company, as well as incorporating those quasi-financial factors which have a direct bearing on the financial flows, such as the placement of purchase orders.

4.4.1 Model Structure

Figure 6 depicts the model structure using the same symbols as for the previous two models, while Figure 7 shows the level element definitions associated with each of the twelve levels incorporated into the model. Five of these levels serve as accumulators of flows, with the flows concerned being the major expenses classifications and the flows required for the determination of gross margin as a net flow. As with the Cash Planning Model, all levels are defined in monetary terms and the model solution interval is set at one month, with a planning horizon of up to sixty months.

The accumulator levels are re-set to zero at twelve month intervals, so that their values constitute the basis for year-to-date income statements. In the course of this zero setting the contents of these levels must be transferred out to the 'Capital' level, via the linkages 11,12,13,14 and 19. In accounting terms this constitutes income appropriation in effect. The accounting significance generally of the levels and flow linkages, is that the former represent account balances, and the latter represent transactions, with a linkage 'head' signifying a debit and a linkage 'tail' signifying a credit. The resultant effect is that debit balances are represented in the system as positive valued levels, and credit balances as negative valued levels.

Ten transformation matrices are defined in the model, all of which perform a switching function with the exception of T3 which carries the resource input proportions associated with each finished product classification. Nine base flow vectors are defined, one for each of the flows in the system which are determined exogenously, either wholly or in part. Seventeen modulation matrices are defined and three delays, the latter to reflect the lag patterns associated with purchase order deliveries, creditor payments, and debtors collections.

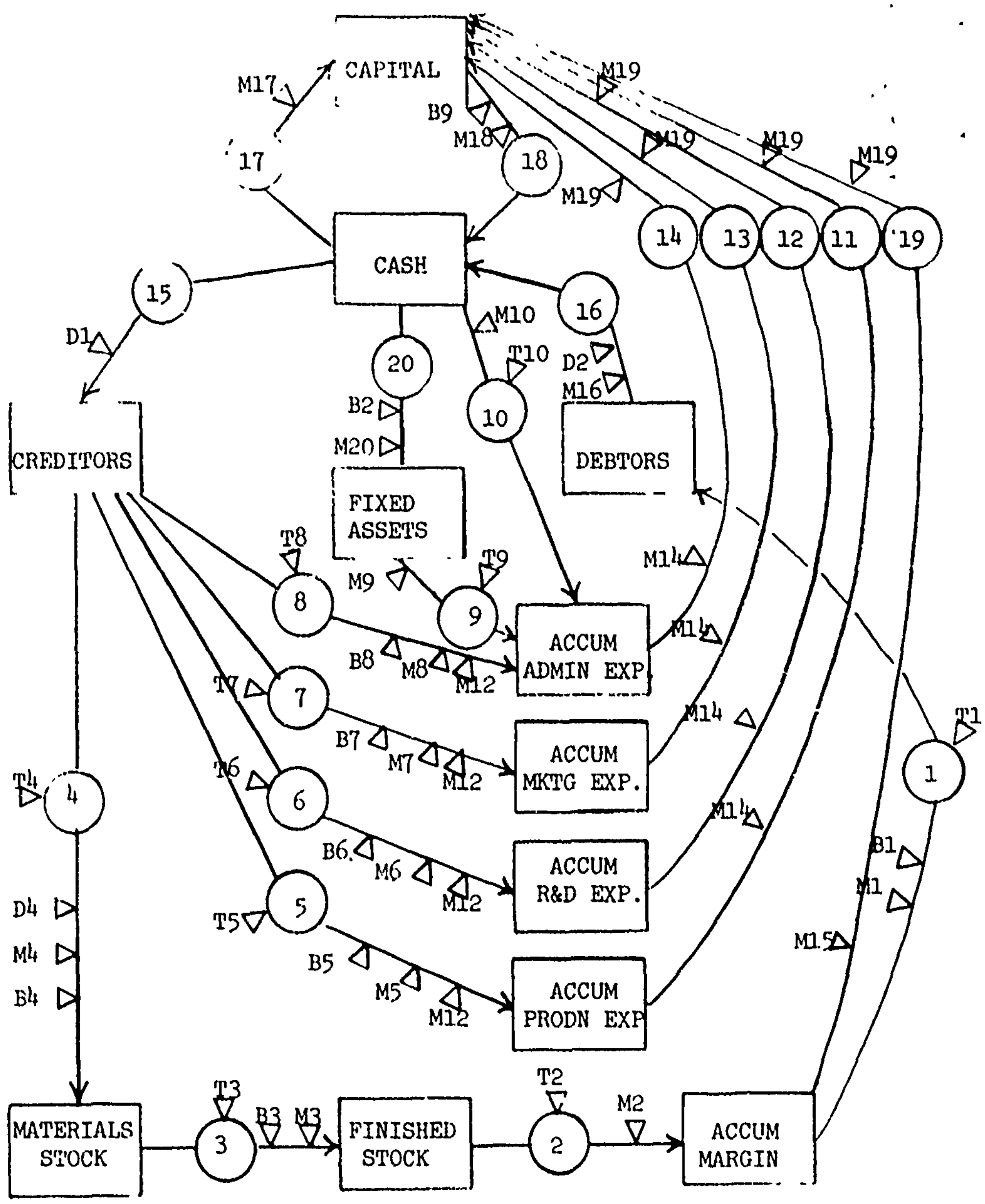


Figure 6. The Corporate Planning Model

CAPITAL	CASH	FIXED ASSETS
Debt Capital	Cash at Bank	Land & Buildings
Equity Capital		Plant & Equipment
Revenue Reserves	CREDITORS	Motor Vehicles
Capital Reserves	Bills Payable	
Prov. for Div.	Local Creditors	DEBTORS
Prov. for Tax	Wages Payable	Local Debtors
		Export Debtors

ACCUM.ADMIN.EXP.	ACCUM.MKTG.EXP.	ACCUM.PRODN.EXP.
Volume Dependent	Volume Dependent	Volume Dependent
Vehicle Dependent	Vehicle Dependent	Vehicle Dependent
Staff Dependent	Staff Dependent	Staff Dependent
Other	Other	Other
Interest	(for Local Mktg.	
Depreciation	& repeated for	
	Export Mktg.)	

ACCUM.R&D EXP.	MATERIALS	FINISHED STOCK
Volume Dependent	Local Materials	Manufactured Products
Vehicle Dependent	Import. Materials	Wholesale Products
Staff Dependent	Wholesale Products	
Other	Direct Wages	

ACCUM. MARGIN
Manufacturing Sales
Wholesale Sales
Export Sales
Manufacturing C.O.S.
Wholesale C.O.S.
Export C.O.S.
Commissions

Figure 7. Level Element Definitions for the Corporate Planning Model

4.4.2 Model Input

The input required for the model is summarised below.

1. Level Initial Values - zero for the five accumulator levels, with the remaining levels being assigned the appropriate opening balance values.

2. Flow Transformation Matrices - as noted above, all of these except T3 perform a switching function and thus have coefficients equal to zero or one. The coefficients of T3 comprise, for each finished product classification, split-up of its total cost into the proportions associated with each resource, or materials classification, (the latter also includes a category for direct labour).

3. Base Flow Vectors - B1, B2, and B4 initialise the flows for sales, production, and purchase order placements respectively. B5, B6, B7, and B8 initialise the flows for those expenses classifications which require it. B2 and B9 relate to the flows associated with capital expenditure, and the raising of new capital.

4. Modulation Matrices - M1, M2, and M4 carry the respective growth patterns for sales, production and purchase order placements, while matrices M5 to M8 carry the real growth patterns for the expenses classifications. M9 carries the depreciation rates for fixed assets, M10 the interest rate for bank overdraft, and M12 provides for the separate specification of the inflationary component of expenses growth. M14 and M15 perform the special function of triggering the transfer of the contents of the accumulator levels to the Capital level at twelve monthly intervals

   The matrices M17, M18, and M20 trigger the timing of tax and dividend payments, borrowing of new capital, and capital expenditure respectively, while M19 carries the coefficients associated with income appropriation and taxation rate.

5. Delay Matrices - D1 is an exponential delay reflecting the payments lag patterns associated with each of the defined creditor classifications. D2 and D4 are both boxcar delays reflecting the lag patterns associated with each of the defined classifications for debtors collections and purchase order deliveries respectively.

6. Cycles Matrix - this comprises the same ten uniquely defined cyclical index series as specified for the Cash Planning Model.

4.4.3 Model Output

Sample output, in tabular form, for one five year run of the model is shown in Appendix IV(a). This report shows only the 'year-end' results, and the expansion of this into monthly results for the immediate year ahead, is shown in Appendix IV(b). Appendix IV(c) shows some graphical output for the same run of the model, displaying the dynamic behaviour of a selection of factors including debtors, stocks, and three key financial ratios.

5. CONCLUSION

The modelling system described in this paper is intended to provide management with a flexible basis for dynamic simulation in the corporate planning context, whereby multiple futures for the organisation, in quantitative terms, can be prepared and updated with the minimum of model development and maintenance effort, and maximum vertical, horizontal, and chronological goal structure integration. In any given application, the system is intended to constitute a central, general purpose modelling facility, to be supported by, or used in conjunction with, special purpose models, (such as those involving the use of optimisation techniques), as the circumstances dictate.

The system in its present form has limitations of both a conceptual and operational nature. The conceptual limitations arise largely from the fact that as a package, it lacks the equation formulation flexibility characteristic of high level simulation languages. For any given model developed under the system, the equation set is derived automatically from a finite library of options, in accordance with the responses tendered in the course of the model specification phase. As far as the more highly structured and mechanistic relationships of an organisation are concerned this does not constitute a problem, but where it is essential to explicitly represent decision processes involving complex information flows, rather than merely reflect them as exogenous influences, the utility of GENSIM will compare unfavourably with a high level simulation language such as DYNAMO.

In brief, the conceptual framework upon which GENSIM is based is one in which the facility to cope with the less structured aspects of the corporate system has been traded off for the facility to cope with disaggregation, and 'structured' complexity in a flexible, and operationally

efficient manner.

The operational limitations of the package are summarised in the following terms:

1. Planning Horizon - not more than 60 periods
2. Number of Level Vectors - not more than 20
3. Number of Linkages - not more than 50
4. Number of Level or Flow Vector Elements - not more than 10
5. Number of Transformation Matrices - not more than 20
6. Number of Modulation Matrices - not more than 20
7. Number of Delayed Flows - not more than 10
8. Number of Cyclical Series - not more than 20

Conversion of the system over to a large main-frame computer would permit the construction of much larger models, and undoubtedly faster processing in the case of any given model. The applications to date however have not revealed any serious practical disadvantages arising from the above constraints.

In conclusion, the extent to which the system achieves its stated aims, and the full practical ramifications of its limitations, have yet to be established. Evaluation should proceed,however with due regard for the comparitive utility in the same circumstances, of the available alternatives.

REFERENCES

(1) Butterworth, J.E. and Sigloch, B.A., 1971; A Generalised Input-Output Model and Some Derived Equivalent Systems, The Accounting Review, November, pp. 700-16.

(2) Cantley, M.F., 1973; Corporate Planning - A Review of Questions and Answers, Omega, Vol. 1, No.1, pp 55-57.

(3) Forrester, Jay W., 1961; Industrial Dynamics, The M.I.T. Press, Cambridge, Massachusetts.

(4) Naylor, T.H., and Mansfield, M.J., 1977; The Design of Computer Based Planning and Modelling Systems, Long Range Planning, Vol. 10, February, pp 16-25.

(5) Perrin, H.F.R., 1971; Focus The Future - An Introduction to Long Range Corporate Planning, Management Publications Ltd., London.

APPENDIX I TERMINAL CONVERSATION FOR DEFINING THE STRUCTURE OF THE HALLMARK CASH PLANNING MODEL

```
RUN GENINP
MODEL NAME IS    OK Y OR N? N
CORRECT NAME ( 6 characters )? ALPHAC
MODEL NAME IS ALPHAC  OK Y OR N? Y
SPECIFICATIONS OR DATA? S
DO YOU WISH TO SET UP NEW FILES Y OR N? Y
STARTING WITH LEVEL SPECS- Y OR N? Y
ZERO ALL FILES Y OR N? Y
NUMBER OF LEVELS? 4
NUMBER OF LINKAGES? 3
ENTER TYPE OF SOLUTION INTERVAL (Y, Q, M12, M13, W, D5, OR D7)? M12
PLANNING HORIZON ? 12

INPUT LEVEL SPECIFICATIONS
FOR LEVEL  1  O.K. ? Y
Description? ACCUMULATED CASH
Dimension? 10
TYPE- 1=Monetary Unbounded 2=Monetary Lower Bound 0 3=Physical Unbounded
4=Physical Lower Bound 0 ? 1
Zero-ising Interval (OR "0")? 1
Level Vector Element Summation procedure (S , P , or N)? S
ELEMENT DESCRIPTIONS Y OR N? N
FOR LEVEL  2  O.K. ? Y
Description? ACCUMULATED PURCHASES
Dimension? 2
TYPE- 1=Monetary Unbounded 2=Monetary Lower Bound 0 3=Physical Unbounded
4=Physical Lower Bound 0 ? 2
Zero-ising Interval (OR "0")? 1
Level Vector Element Summation procedure (S , P , or N)? S
ELEMENT DESCRIPTIONS Y OR N? N
FOR LEVEL  3  O.K. ? Y
Description? ACCUM. OTHER PAYMENTS
Dimension? 7
TYPE- 1=Monetary Unbounded 2=Monetary Lower Bound 0 3=Physical Unbounded
4=Physical Lower Bound 0 ? 2
Zero-ising Interval (OR "0")? 1
Level Vector Element Summation procedure (S , P , or N)? S
ELEMENT DESCRIPTIONS Y OR N? N
FOR LEVEL  4  O.K. ? Y
Description? ACCUMULATED SALES
Dimension? 2
TYPE- 1=Monetary Unbounded 2=Monetary Lower Bound 0 3=Physical Unbounded
4=Physical Lower Bound 0 ? 2
Zero-ising Interval (OR "0")? 1
Level Vector Element Summation procedure (S , P , or N)? S
ELEMENT DESCRIPTIONS Y OR N? N
```

APPENDIX I Continued...

```
INPUT SPECIFICATIONS FOR LINKAGE  1 O.K.? Y
LINKAGE TYPE (IC, II, IX, IF, SC, SI, SX, SF, GC, GI, GX, GR GF)? GI

INPUT INFLOW DETAILS FOR LINKAGE  1
What Level does it flow into ? 2
Inflow Dependency Type (E, L, I, X, or M)? E
Determinant Flow Number? 1
INFLOW PATTERN (I,E OR B)? I
Modulating Tag No.? 1
Cost Vector Dependency (E OR "CR")?
Transformation Matrix Tag No.? 1
INPUT OUTFLOW DETAILS FOR LINKAGE 1
What level does it flow from? 1
Outflow Dependency (E, L, I, X, or M)? I
Determinant Flow No.? 1
OUTFLOW PATTERN (I,E OR B)? B
Boxcar delay tag no? 1
Cost Vector Dependency (E OR "CR")?

INPUT SPECIFICATIONS FOR LINKAGE  2 O.K.? Y
LINKAGE TYPE (IC, II, IX, IF, SC, SI, SX, SF, GC, GI, GX, OR GF)? GI

INPUT INFLOW DETAILS FOR LINKAGE  2
What Level does it flow into ? 3
Inflow Dependency Type (E, L, I, X, or M)? E
Determinant Flow Number? 2
INFLOW PATTERN (I,E OR B)? I
Modulating Tag No.? 2
Cost Vector Dependency (E OR "CR")?
Transformation Matrix Tag No.? 2
INPUT OUTFLOW DETAILS FOR LINKAGE 2
What level does it flow from? 1
Outflow Dependency (E, L, I, X, or M)? I
Determinant Flow No.? 2
OUTFLOW PATTERN (I,E OR B)? I
Cost Vector Dependency (E OR "CR")?
```

APPENDIX I Continued...

```
INPUT SPECIFICATIONS FOR LINKAGE  3 O.K.? Y
LINKAGE TYPE (IC, II, IX, IF, SC, SI, SX, SF, GC, GI, GX, OR GF)? GX

INPUT INFLOW DETAILS FOR LINKAGE  3
What Level does it flow into ? 1
Inflow Dependency Type (E, L, I, X, or M)? X
Determinant Flow Number?  3
 INFLOW PATTERN (I,E OR B)? B
 Boxcar delay tag no? 3
 Cost Vector Dependency (E OR "CR")?
 Transformation Matrix Tag No.? 3
 INPUT OUTFLOW DETAILS FOR LINKAGE 3
 What level does it flow from? 4
 Outflow Dependency (E, L, I, X, or M)? E
 Determinant Flow No.? 3
 OUTFLOW PATTERN (I,E OR B)? I
 Modulating Tag No.? 3
 Cost Vector Dependency (E OR "CR")?
 DO YOU WISH TO ENTER DATA? N
```

APPENDIX II

GENERALISED SIMULATION SYSTEM - VERSION 2 *** HALLMARK INDUSTRIES CASH PLANNING MODEL
RUN MONTH 1=OCTOBER
RESULTS FOR MONTH ENDING:-

REF	DESCRIPTION	UNITS	0	1	2	3	4	5	6	7	8	9	10	11	12
101	COLLECTIONS		0	298	235	188	175	204	214	221	233	262	315	358	391
102	PURCHASES PAYMEN.		0	-132	-132	-135	-137	-138	-148	-152	-164	-164	-164	-171	-171
103	MFG. WAGES		0	-37	-35	-37	-35	-57	-60	-63	-66	-63	-57	-60	-60
104	OVERHEADS		0	-53	-48	-50	-50	-62	-65	-65	-68	-71	-76	-80	-82
105	NET CASH FROM OPS		0	76	19	-35	-47	-54	-59	-60	-65	-36	18	47	78
106	CAPITAL EXPEND.		0	-22	0	-11	-11	0	0	-33	-51	-52	-22	-22	-11
107	REPAYMENTS/BORROWINGS		0	0	0	0	0	0	0	0	0	0	0	0	0
108	TAX PAYMENTS		0	0	0	0	-6	0	0	0	0	0	-5	0	0
109	DISTRIBUTIONS		0	50	0	0	0	0	0	0	0	0	0	0	0
110	NET CASH BALANCE		-57	47	19	-46	-64	-54	-59	-93	-115	-88	-9	25	[illegible]
111	SALES		0	210	130	171	181	244	210	230	253	313	402	391	428
112	PURCH. ORDERS PLACED		0	171	171	171	171	171	171	171	171	171	171	257	257

APPENDIX III

GENERALISED SIMULATION SYSTEM - VERSION 2 *** HALLMARK PRODN PLANNING MODEL

RUN YEAR 1=1978

RESULTS FOR YEAR ENDING:-

REF	DESCRIPTION	UNITS	0	1	2	3	4	5
201	EL. 1 PACK NYLON	COST P/U	1.200	1.200	1.320	1.452	1.597	1.757
201	EL. 2 TENT NYLON	COST P/U	1.850	1.850	2.035	2.239	2.462	2.709
201	EL. 3 SB FILL	COST P/U	3.500	3.500	3.850	4.235	4.659	5.124
201	EL. 4 WEBBING	COST P/U	0.450	0.450	0.495	0.545	0.599	0.659
201	EL. 5 METAL TUBING	COST P/U	1.200	1.200	1.320	1.452	1.597	1.757
201	EL. 6 MISCELLANEOUS	COST P/U	0.250	0.250	0.275	0.303	0.333	0.366
201	EL. 7 METALWORK M/HR	COST P/U	2.500	2.500	2.750	3.025	3.328	3.660
201	EL. 8 PACK MFG. M/H	COST P/U	2.800	2.800	3.080	3.388	3.727	4.099
201	EL. 9 OTHER MFG. M/H	COST P/U	2.800	2.800	3.080	3.388	3.727	4.099
202	EL. 1 P1 PACK FMS	COST P/U	8.360	8.361	9.197	10.117	11.129	12.242
202	EL. 2 P2 PACK FMS	COST P/U	7.990	7.997	8.797	9.677	10.644	11.709
202	EL. 3 PACK NYLON	COST P/U	1.200	1.200	1.320	1.452	1.597	1.757
202	EL. 4 WEBBING	COST P/U	0.450	0.450	0.495	0.545	0.599	0.659
202	EL. 5 MISCELLANEOUS	COST P/U	0.250	0.250	0.275	0.303	0.333	0.366
202	EL. 6 PACK MFG M/HRS	COST P/U	2.800	2.800	3.080	3.388	3.727	4.099
203	EL. 1 P1 PACKS	COST P/U	12.550	19.873	19.874	21.861	24.047	26.452
203	EL. 2 P2 PACKS	COST P/U	10.250	18.774	18.781	20.659	22.725	24.998
203	EL. 3 T1 TENTS	COST P/U	21.050	21.054	23.159	25.475	28.023	30.825
203	EL. 4 T2 TENTS	COST P/U	32.360	32.358	35.593	39.153	43.068	47.375
203	EL. 5 T3 TENTS	COST P/U	45.020	45.018	49.519	54.471	59.918	65.910
203	EL. 6 SB1 SLPG BAGS	COST P/U	13.180	13.183	14.501	15.951	17.546	19.301
203	EL. 7 SB2 SLPG BAGS	COST P/U	12.610	12.610	13.871	15.258	16.784	18.463
203	EL. 8 ACCESSORIES	COST P/U	4.350	4.355	4.791	5.270	5.797	6.376

GENERALISED SIMULATION SYSTEM - VERSION 2 *** HALLMARK PRODN PLANNING MODEL

RUN YEAR 1=1978

RESULTS FOR YEAR ENDING:-

REF	DESCRIPTION	UNITS	0	1	2	3	4	5
3	EL. 1 P1 PACKS	PROD. UNIT	0	2600	3700	5013	6482	7480
3	EL. 2 P2 PACKS	PROD. UNIT	0	1850	2633	3567	4612	5322
3	EL. 3 T1 TENTS	PROD. UNIT	0	865	1259	1706	2206	2546
3	EL. 4 T2 TENTS	PROD. UNIT	0	675	961	1301	1683	1942
3	EL. 5 T3 TENTS	PROD. UNIT	0	250	356	482	623	719
3	EL. 6 SB1 SLPG BAGS	PROD. UNIT	0	145	206	280	361	417
3	EL. 7 SB2 SLPG BAGS	PROD. UNIT	0	325	462	627	810	935
3	EL. 8 ACCESSORIES	PROD. UNIT	0	3450	4909	6652	8601	9926
	TOTALS FOR ACCUM. PRODN.		0	159163	236867	353021	502121	637409
101	PACK NYLON		0	6863	9765	13231	17108	19743
102	TENT NYLON		0	15115	21509	29142	37682	43486
103	SB FILL		0	746	1061	1438	1859	2145
104	WEBBING		0	3491	4967	6730	8702	10042
105	METAL TUBING		0	8160	11611	15732	20342	23475
106	MISCELLANEOUS		0	66462	94575	128138	165689	191210
107	METALWORK	MAN HRS	0	10369	14754	19990	25849	29830
108	PACKS	MAN HRS	0	13039	18554	25138	32505	37512
109	TENTS & OTHER	MAN HRS	0	10702	15229	20634	26680	30790

APPENDIX IV

GENERALISED SIMULATION SYSTEM - VERSION 2 *** HALLMARK INDUSTRIES CORPORATE PLANNING MODEL

RUN MONTH 1=JANUARY 1978

RESULTS FOR MONTH ENDING:-

REF	DESCRIPTION	UNITS	0	12	24	36	48	60
101	TOTAL CAPITAL		-740	-724	-871	-1093	-1435	-1939
118	CREDITORS		-181	-311	-450	-597	-736	-871
121	ACCUM. PROFIT		0	-196	-246	-361	-519	-599
102	CASH		-113	-115	52	-148	164	598
103	TOTAL LIABILITIES	$'000	-1034	-1347	-1515	-2198	-2526	-2811
119	FIXED ASSETS		404	589	671	1164	1318	1255
122	FINISHED STOCK (MFG.)		193	240	285	343	351	410
123	WHOLESALE STOCK		24	13	19	39	66	109
124	MATERIALS STOCK		307	277	251	275	319	489
120	DEBTORS		106	228	288	377	471	547
104	TOTAL ASSETS	$'000	1034	1347	1515	2198	2526	2811

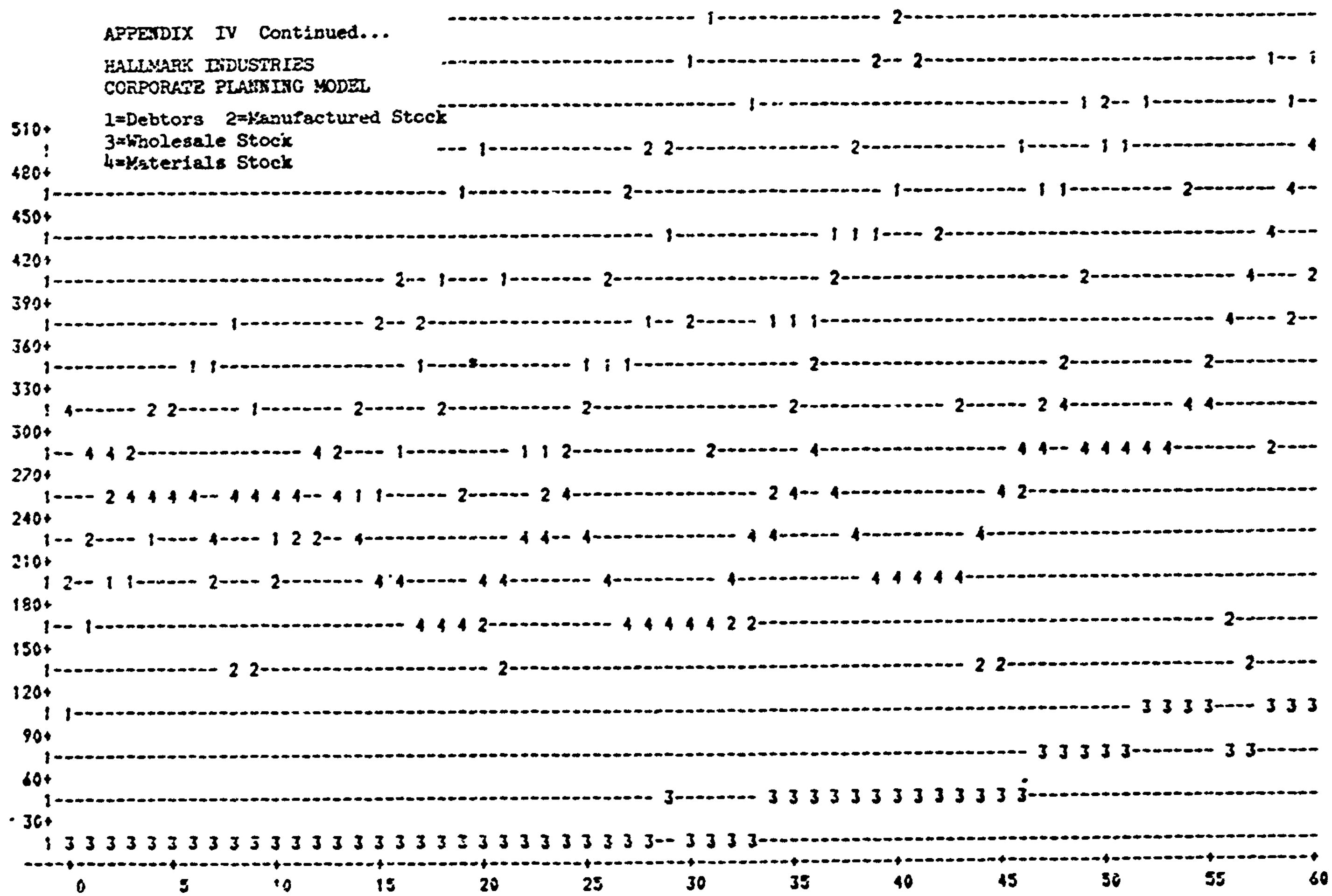
125 LOCAL MFG. SALES		0	1680	2123	2768	3462	4013
126 WHOLESALE SALES		0	168	212	277	346	401
127 EXPORT SALES		0	600	1050	1500	2026	2326
128 TOTAL SALES		0	2448	3385	4545	5834	6740
129 LOCAL MFG. MARGIN		0	521	658	858	1073	1244
130 WHOLESALE MARGIN		0	45	57	75	93	108
131 EXPORT MARGIN		0	90	158	225	304	349
132 TOTAL MARGIN	$'000	0	656	873	1158	1471	1701
133 ADMIN. O/H		0	103	124	133	160	179
134 LOCAL MKTG. O/H		0	141	171	213	264	316
135 EXPORT MKTG. O/H		0	69	152	230	256	286
136 RESEARCH & DEV. O/H		0	54	59	65	71	78
137 PRODUCTION O/H		0	112	139	175	217	260
138 TOTAL OVERHEAD		0	478	645	815	969	1120
140 COMMISSIONS		0	18	18	18	18	18
139 NET PROFIT	$'000	0	196	246	361	519	599

APPENDIX IV Continued...

116 RETURN ON EQUITY %	0.000	28.978	29.466	32.141	33.483	29.516
117 RETURN ON SALES %	0.000	8.017	7.268	7.932	8.901	8.894
113 LIQUID RATIO	0.586	0.732	0.642	0.631	0.641	0.628
114 EQUITY RATIO	0.465	0.439	0.474	0.438	0.509	0.595
115 CURRENT RATIO	1.944	1.720	1.872	1.220	1.619	2.823
110 LOCAL GROSS MARGIN %	0.000	31.000	31.000	31.000	31.000	31.000
111 WHOLESALE GROSS MARGIN %	0.000	27.000	27.000	27.000	27.000	27.000
112 EXPORT GROSS MARGIN %	0.000	15.000	15.000	15.000	15.000	15.000
109 TOTAL MARGIN %	0.000	26.804	25.786	25.476	25.207	25.241
105 ADMIN OVERHEAD %	0.000	4.206	3.653	2.925	2.751	2.653
106 MARKETING OVERHEAD %	0.000	8.574	9.550	9.743	8.922	8.934
107 R&D OVERHEAD %	0.000	2.186	1.739	1.425	1.221	1.163
108 PRODUCTION OVERHEAD %	0.000	4.557	4.109	3.847	3.722	3.865

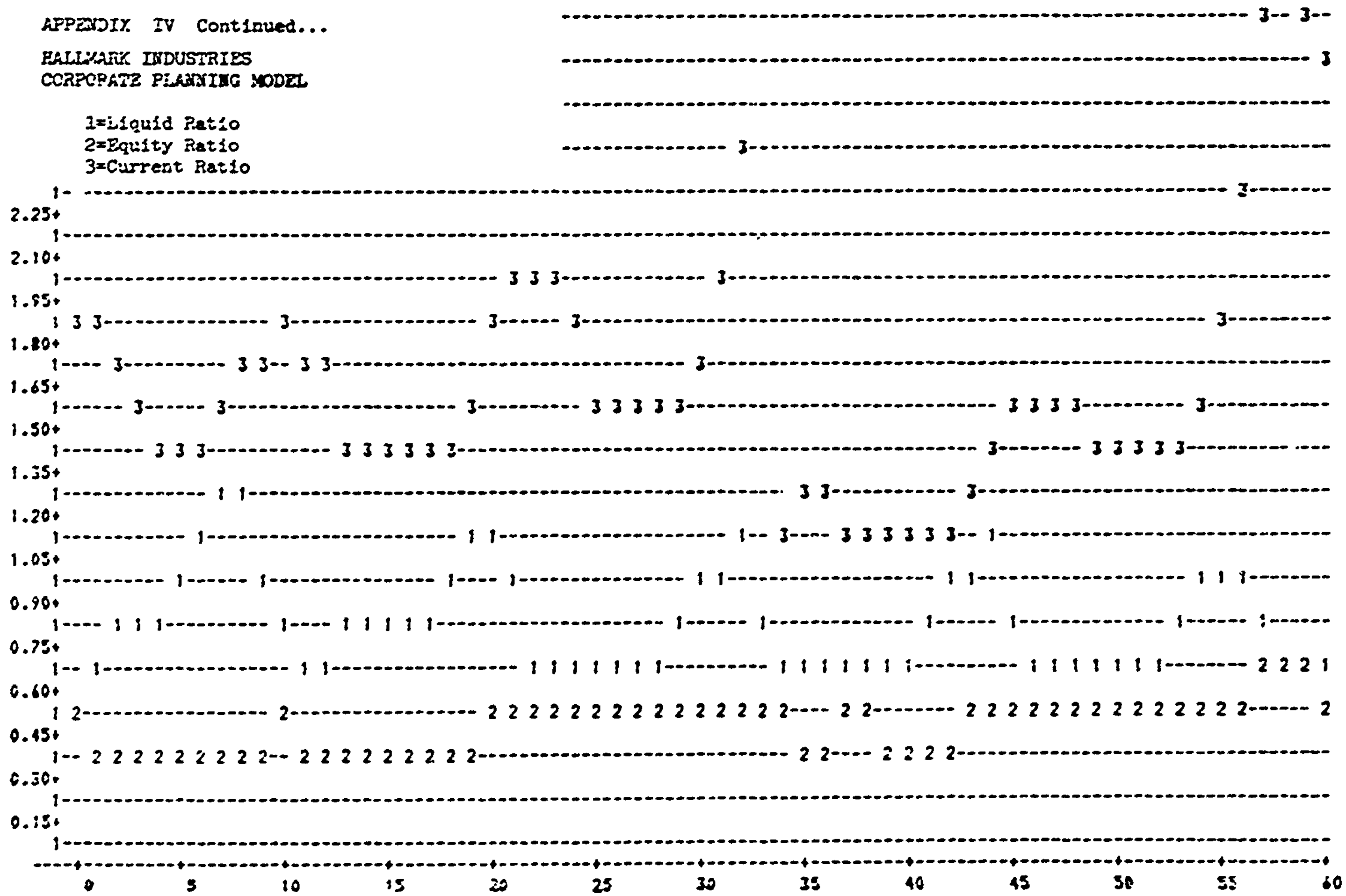
APPENDIX IV Continued...

HALLMARK INDUSTRIES
CORPORATE PLANNING MODEL



APPENDIX IV Continued...

HALLMARK INDUSTRIES
CORPORATE PLANNING MODEL



OPTIMAL PRODUCTION PLANNING FOR A PENCIL INDUSTRY - A CASE STUDY

VORATAS KACHITVICHYANUKUL[1] and M. NAWAZ SHARIF[2]

Industrial Engineering and Management Division
Asian Institute of Technology
Bangkok, Thailand

ABSTRACT. The paper presents a production plan for an existing pencil industry and its expansion plan to meet anticipated future demand at a minimum cost. A linear programming model was used for optimization. The demand was forecasted by multiple linear regression analysis, using past sales and other economic parameters. Based upon the capacities of different machines in the existing plant, two expansion plans were proposed. The incremental fixed costs, maximum capacities, and unit costs of production were evaluated. The optimization model was then run for many cases: one shift, two shifts, three shifts operation, and one shift plus overtime operation for the existing plant, expansion plan 1, and expansion plan 2. The results were compared to get the alternative which gives minimum cost over the study period.

---

[1] Research Associate

[2] Associate Professor of Industrial and Systems Engineering

1. INTRODUCTION

In Thailand, no pencil industry was established until late 1973. The main reason was that the imported pencils were very cheap and the government did not provide any incentive for local investment. The pencil industry under study has been in existence for the last five years (1974-1978). The machines and techniques used in the industry are relatively simple and need less technical knowhow. The main problem in the industry being studied is to increase its profitability through proper production planning. At the time of studying, management of the firm was interested in finding answers to the following questions:

a) How to schedule the production in order to meet the market demand in an optimal manner.

b) With the expectation of the growth in sales, the next problem is to determine when to expand the existing facilities and to what extent.

2. FORMULATION OF THE MODEL

2.1. This section presents a linear programming model for determining the most economical way of production subjected to a given definite seasonal demand. From the forecasted demand the monthly sales of the firm for next few years can be estimated and is taken as given for this optimization problem. The model is developed for a given plant capacity and a given planning period, but it can be applied for several planning periods and for different plant capacities as well.

For a given plant capacity, the output of the plant depends on the intensity with which the equipment is used, the size of the workforce, the number of shifts and the level of seasonal inventory stocks.

The planning horizon of the firm can be divided into several planning periods, each of which is one year. For each planning period, we have to schedule the workforce in such a way that the total cost over the study period will be minimized.

The basic decision variables are the monthly production, the inventory levels needed to meet the monthly demand and the shortage amounts if the demand cannot be met. The production in any month can consist of several choices:

$x(t)$, normal shift production with unit cost C_x;

$y(t)$, second shift production with unit cost C_y;

$z(t)$, Third shift production with unit cost C_z;

where t is the time subscript for month 1 to month 12 within each planning period ($t=1, 2, \ldots, 12$). The unit cost of production excludes the raw material costs since they are independent of the workforce level.

2.2. Constraints

2.2.1. Capacity Constraints:

The output produced from each of the production types for each level of plant capacity is limited.

$$x(t) \leq X \qquad t=1, 2, \ldots, 12$$
$$y(t) \leq Y \qquad t=1, 2, \ldots, 12$$
$$z(t) \leq Z \qquad t=1, 2, \ldots, 12$$

where X, Y and Z are the maximum capacity of first, second and third shift production respectively and for a specific level of plant capacity.

2.2.2. Inventory Constraints:

The number of units held in inventory at the end of each month should be limited due to storage limitations.

$$I(t) \leq I \qquad t=1, 2, \ldots, 12$$

where

$I(t)$: Stock of inventory at the end of month t.

I : Maximum number of units that can be held in inventory.

In view of the assumption that the demand pattern is periodic over twelve months, the inventory level at the end of the planning period should equal to the inventory level at the beginning of the planning period multiplied by the secular growth rate G.

$$I(12) = (1+G)I(0)$$

where

$I(0)$: Number of units held in inventory at the beginning of month 1. The value of $I(0)$ should be given as the input to the model and will be

expressed as fraction of the maximum storage capacity.

2.2.3. Demand Constraints:

It is desirable that the production plan must meet the forecasted demand for each month.

$$x(t) + y(t) + z(t) + I(t-1) - I(t) + S(t) = D(t)$$

where

$t=1, 2, \ldots, 12$

$S(t)$: Amount of shortage during month t.

$D(t)$: Demand during month t.

2.3. The Objective Function

The optimal production plan must satisfy all the preceding equations as well as minimizing the sum of ·roduction cost, inventory holding cost and penalty cost of not meeting the demand. The objective function can be written in the following form:

$$\sum_{t=1}^{12} [C_x x(t) + C_y y(t) + C_z z(t) + C_i I(t) + C_s S(t)]$$

where

C_i: Inventory holding cost per unit per month.

C_s: Shortage penalty cost per unit.

3. EVALUATION OF INPUT PARAMETERS

The parameters which are needed as input to the model and should be evaluated are

1. Monthly demand of the product.
2. Inventory storage capacity.
3. Production capacity limits.
4. Production cost.
5. Inventory holding cost,
6. Shortage penalty cost.

3.1. Estimation of Future Demand

In Thailand, there were no local pencil producer until late 1973. All the pencils were imported and the amounts imported can be used to estimate the demand of the market

all over Thailand. The past records of imported pencils are correlated with four relevant factors - total population, number of school children, per capita income and time. The regression method used is multiple linear regression from IBM Scientific Subroutine package.

By comparing the correlation coefficients and the standard errors of estimate, it was found out that the past data reflect a definite trend if correlated with the following factors; total population of Thailand, number of school children and time.

Based upon past sales record of the firm, the market share of the firm is estimated to be twenty five percents of the overall market. The yearly demand of the firm for next four years is then estimated and shown in Table 3.1.

Table 3.1 Forecasted Demand of the Firm for Next Four Year

YEAR	DEMAND IN UNITS
1977	182,250
1978	189,084
1979	196,000
1980	203,084

The study of the past records of imported pencil indicates that there is a definite seasonal variation in the market demand of pencil in Thailand. The source of past records used is named in [6].

The monthly demands are expressed as fractions of yearly demand as shown in column 2, Table 3.2. By assuming that the demand pattern of the firm is the same as the overall market, the monthly demands of the firm can be calculated and are summarized in Table 3.2.

3.2. Estimation of Storage Capacity

The maximum storage capacity can be estimated by considering the floor space available for finished product inventory in the existing layout, the crushing strength and the dimension of the box.

Table 3.2 Forecasted Monthly Demand for the Firm in Units of Pencil, Based Upon 25% of the Overall Market

MONTH	FRACTION	1977	1978	1979	1980
Jan	0.069	12,575	13,047	13,524	14,013
Feb	0.081	14,762	15,316	15,876	16,450
Mar	0.161	29,342	30,442	31,556	32,696
Apr	0.141	25,697	26,661	27,636	28,635
May	0.138	25,151	26,093	27,048	28,025
Jun	0.087	15,856	16,450	17,052	17,668
Jul	0.070	12,758	13,236	13,720	14,216
Aug	0.061	11,117	11,534	11,956	12,338
Sep	0.043	7,837	8,131	8,428	8,733
Oct	0.031	5,650	5,862	6,076	6,296
Nov	0.057	10,388	10,778	11,172	11,576
Dec	0.061	11,117	11,534	11,956	12,338
TOTAL	1.000	182,250	189,084	196,000	203,084

3.3. Estimation of Production Capacity Limits

For any production system, the capacity is limited by the machine with the slowest rate. The rates of different machines and/or operating units of the existing system are summarized in the first column of Table 3.3 and it appears that the machine with the slowest rate is piercing machine with the hourly rate of 65 units. The existing system operates on an eight-hour-shift basis with about seven hours of actual operating time (the firm allows one hour for lunch plus personal allowance and operates 25 days a month). Based upon the figures, the output per shift and the output per shift per month can be calculated.

In order to cope with the increasing demand, two possible expansion plans were proposed. These two plans offer two alternatives to increase the production capacity through the addition of machines and/or operating units. The details of

these two expansion plans and the resultant values such as production capacity, incremental investment cost, increase in labors, and horsepower requirements are summarized in the second and third column of Table 3.3.

Table 3.3 Summary of Capacities, Costs, Power and Labor Requirements of Three Different Alternatives

OPERATION NAME	UNITS/HOUR FOR ALTERNATIVE *		
	1	2	3
Width Cutting	82.5	165.0	165.0
Grooving	90.0	135.0	135.0
Lead Filling	90.0	112.5	123.75
Rough End Cutting	187.5	187.5	187.5
Shaping	90.0	135.0	135.0
Painting	72.6	108.9	108.9
Fine End Cutting	143.3	143.3	143.3
Hot Foil Stamping	115.0	115.0	115.0
Shouldering	70.8	106.2	106.2
Rubber Inserting	69.2	103.7	103.7
Piercing	65.0	97.5	97.5
Total HP required	34.0	43.0	41.0
Workers per Shift	49	56	65
Incremental Cost (total in Baht)	0	312,500	113,500
Incremental Cost (annual in Baht)	0	62,265	22,615
MAXIMUM OUTPUT			
in units per shift	455	682.5	682.5
in units per shift per month	11,375	17,062.5	17,062.5

* Alternative 1: Existing Plant

Alternative 2: Expansion Plan 1

Alternative 3: Expansion Plan 2

3.4. Estimation of Production Costs

3.4.1. Labor Cost

For the existing system, the total number of workers per shift is shown in Table 3.3. The minimum wages set up by the Thai government can be used for the calculation. Therefore, the direct labor cost for the first shift of the existing system can be calculated.

3.4.2. Power Cost

For first shift operation, power will be consumed mainly by the machines. Total cost per shift was estimated using the total horsepowers required in the existing plant.

The indirect costs such as security guards, office workers, etc., are fixed and excluded in this study.

Since the raw material cost was independent of the workforce level, it was excluded and the total cost of production was considered to be the sum of the labor cost and power cost.

For first shift operation, the maximum output for the existing system is 455 units per shift, but for some practical reasons, an efficiency of 90 percents is assumed and the output of the existing system is approximately 400 units per shift (this figure also confirmed with the past performance of the firm). Therefore, the unit cost of production can be estimate.

For the second shift and third shift operations, additional power cost for illumination and additional transportation cost should be taken into consideration. But the values calculated from actual data show that there is no significant differences among the unit costs of these three types of production. In order to make use of the first shift production as much as possible, the unit costs of production of the second shift and the third shift should be high enough to make the linear programming model work effectively. As the unit production cost used here does not include the raw material cost or other fixed cost and more than 90 percents of this cost is labor cost, which will increase for the second and third shifts operation, the unit costs of production for the second shift and the third shift are taken as the unit cost of production of the first shift times 1.5 and 2.5 respectively.

For expansion plan 1 and expansion plan 2, the costs can be calculated in a similar way and all results are presented in Table 3.4.

3.5. Estimation of Inventory Holding Cost

The inventory holding cost of pencils is being considered as the cost of holding the amount of money equivalent to the selling price of those pencils. The approximated interest rate is assumed to be one percent per month and the results are shown in Table 3.4.

Table 3.4 Values of Estimated Cost (Baht)

COST ELEMENT	ALTERNATIVE 1	ALTERNATIVE 2	ALTERNATIVE 3
Production cost (first shift)	3.410	2.553	2.926
Production cost (second shift)	5.115	3.830	4.389
Production cost (third shift)	8.525	6.383	7.315
Inventory holding cost	0.642	0.642	0.642
Shortage penalty cost	10.000	10.000	10.000
Incremental cost (total)	0.000	312,500	113,500
Incremental cost (annual)	0.000	62,265	22,615

3.6. Estimation of Shortage Penalty Cost

The shortage penalty cost is being considered here as the opportunity lost of the profit that the firm should get if there were enough pencils for sales. Therefore, the shortage penalty cost should equal the selling price of the pencil less the total cost of producing the pencil.

For this study, the shortage penalty cost will be taken as 10 Baht per unit.

In the use of the proposed model, the value of the initial inventory at the beginning of month 1 is taken as unknown parameter and will be varied based upon past experiences, subjective judgement and also the results of the model itself.

4. ANALYSIS OF RESULTS

The proposed model was run for three alternatives and four consecutive years by IBM-370/145 system. The value of G is kept to be zero and the value of I(0) is varied. This implied that the operating plans will force the inventory at the beginning and the end of the year to be the same. The total costs over the whole study period (1977 to 1980) are summarized and plotted in Fig. 4.1, Fig. 4.2 and Fig. 4.3 for the cases that allow one shift, two shifts and three shifts operation respectively. All the results show that alternative 2 (expansion plan number 1) is the one which will give the minimum total cost over the whole study period.

For one shift operation, the production capacities cannot keep up with the market demand and the optimal point occurs at higher inventory level.

For the case that allows two shifts production, the capacities are doubled and are very much higher than the market demand, hence the operating plan will never go beyond the second shift and the optimal policy is to maintain lower inventory level.

As the proposed model did not take the idle capacity cost into considerations, the results of the cases that allow two shifts and the cases that allow three shifts are expected to be the same, which can be seen in Fig. 4.2 and Fig. 4.3 that the two sets of curves are almost identical.

The cases where the values of G are not zero are also tried but the total costs are very much higher and are discarded (see Table 4.1).

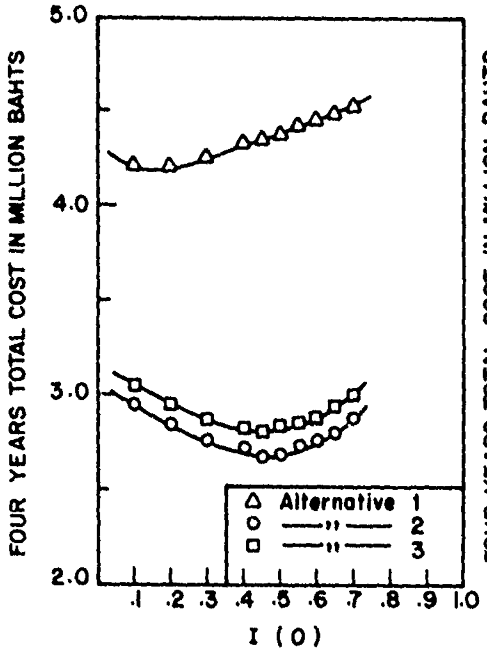


Fig 4.1 Results of the operating plan, one shift operation, G=0

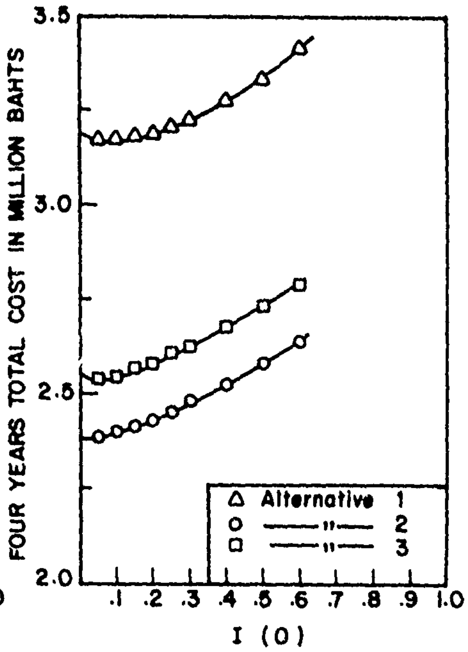


Fig.4.2 Results of the operating plan, two shifts operation, G = 0

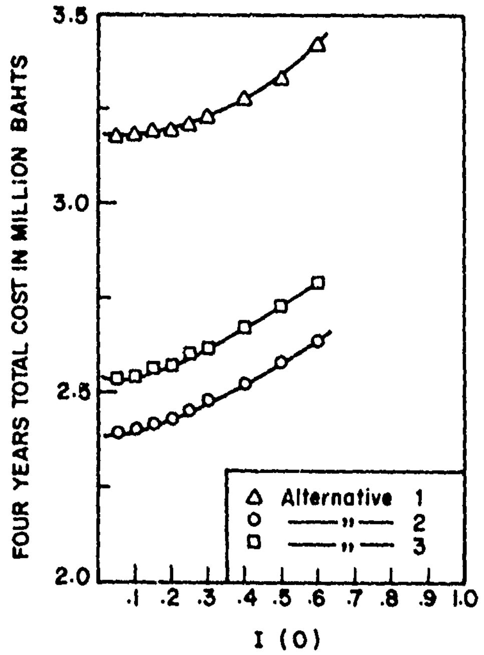


Fig. 4.3 Results of the operating plan, three shift operation, G = 0

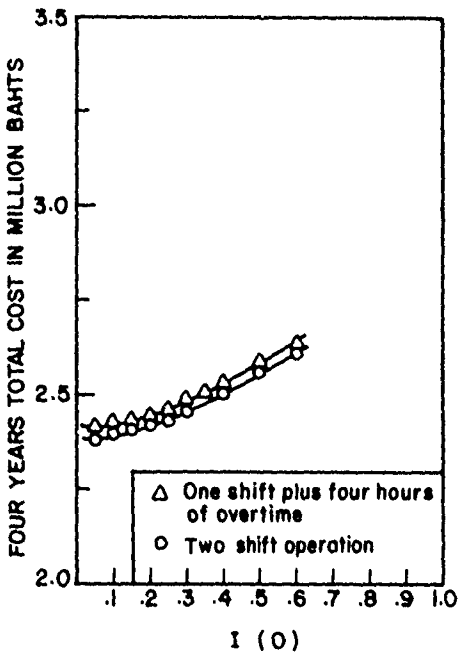


Fig. 4.4 Results of the operating plan, alternative 2, G = 0

Table 4.1 Results of the Operating Plan for One Shift Operation, G $\neq$ 0 (Four Years Total Cost in Million Baht)

G	I(O)	ALTERNATIVE 1	ALTERNATIVE 2	ALTERNATIVE 3
0.40	0.20	4.670	2.960	3.088
0.20	0.35	4.605	2.826	2.960
0.15	0.40	4.574	2.802	2.936
0.10	0.45	4.529	2.775	2.908

From the results of the plan in Table 4.2, it is not practical that the second shift is operated only during some of the month.

Table 4.2 Result of Production Plan for Year 1977 Alternative Number 2

MONTH	1ST SHIFT AMOUNT	2ND SHIFT AMOUNT	INVENTORY AMOUNT	DEMAND
1	6550.0	0.0	0.0	12575.0
2	17062.5	0.0	2300.5	14762.0
3	17062.5	9979.0	0.0	29342.0
4	17062.5	8634.5	0.0	25697.0
5	17062.5	8088.5	0.0	25151.0
6	15856.0	0.0	0.0	15856.0
7	12758.0	0.0	0.0	12758.0
8	11117.0	0.0	0.0	11117.0
9	7837.0	0.0	0.0	7837.0
10	5650.0	0.0	0.0	5650.0
11	10467.5	0.0	79.5	10386.0
12	17062.5	0.0	8025.0	11117.0

To make the plan more realistic, the model was run again by changing the second shift operation to be four hours of overtime production and the results are given in Fig. 4.4 (the model was rerun only for alternative 2 since the other two alternatives gave very much higher cost and were discarded to save the computer time).

The results of the modified plan give slightly higher costs than the case of two shifts operation. But if the idle capacity cost of the second shift is taken into consideration, then the cost of allowing two shifts operation will be very much higher than the cost of overtime operation. Furthermore, it is more practical to operate one shift throughout the year and allow overtime during some of the months than to operate two shifts some of the months and allow idle workforce during most of the time.

The case where shortage cost is equal to 5 Baht per unit is also considered and the results are presented in Fig. 4.5 for alternative 2, one shift operation.

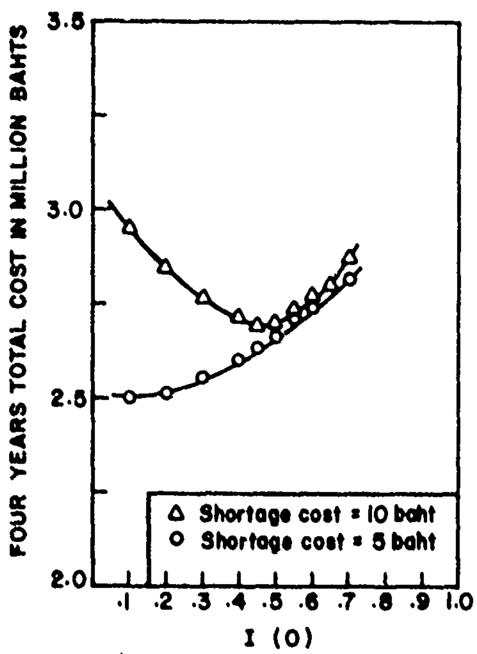


Fig. 4.5 Comparison of the results, alternative 2, G = O

The model gives quite different results compared to the original case with shortage cost equals 10 Baht per unit. The optimal policy will decrease from I(0) = 0.45 to I(0) = 0.1 as the shortage penalty cost reduced from 10 Baht per unit to 5 Baht per unit.

It is also shown that lowering the shortage penalty cost will result in allowing more shortage amount during the years which violates the policy of the firm and is undesirable.

5. SUMMARY

It is important to realize that the study emphasized more on production planning and future expansion of the firm for four years period. It was not intended to cover a detail study of the operating procedures, raw materials and in-process inventory. However, some indication has been given of the possibility of applying the linear programming model to aid the management of small scale industry toward better decision making.

6. REFERENCES

[1] Chang, R.H. and Jones, C.M. (1970), PRODUCTION AND WORK-FORCE SCHEDULING EXTENSIONS, AIIE Transactions, Vol. 2, No. 4, December 1970.

[2] Elwood, S.B. and William, H.T. (1972), PRODUCTION-INVENTORY SYSTEMS PLANNING AND CONTROL, Richard D. Irwin Inc.

[3] Follie, G.M. (1974), A SIMPLE MODEL FOR COMPUTING THE TRADE-OFFS BETWEEN PLANT CAPACITY AND OPERATING STRATEGIES, Operational Research Quarterly, Vol. 25, No. 1, March 1974.

[4] FOREIGN TRADE STATISTICS OF THAILAND, Department of Customs, Bangkok, Thailand

[5] IBM-360 & IBM-370 SUBROUTINE LIBRARY-MATHEMATICS (SL-Math) User's Guide, pp. 279-285.

[6] Kachitvichyanukul, V. (1976), OPTIMAL PRODUCTION PLANNING FOR A PENCIL FACTORY IN THAILAND, Master's Thesis, Asian Institute of Technology, Bangkok, Thailand.

[7] STATISTICAL YEARBOOK OF THAILAND, National Statistical Office, Office of the Prime Minister, Bangkok, Thailand.

MANPOWER MANAGEMENT SYSTEM IN DESIGN OFFICE

WILLIAM F. SHEI

Standard and Code Division,
General Engineering Dept.
China Technical Consultants, Inc.
Taipei, Taiwan, R.O.C.

ABSTRACT. The purpose of this management system is to provide the management levels with accurately and timely reports which can let them to know what the design office going on, and also can let them make decisions in manpower management. Besides, we use reports of this system as the project progress and manpower consumption control. And further more, it will be used as the manpower estimate basis of the future proposals for engineering project.

This system actually is a manpower subsystem of management information system in an engineering company, so it is limited to the design office manpower management only.

This manpower management system uses the CDC Cyber 74 as the tool of analysis and statistics of the design office. First we design a system model as the basis of this system. Then we set a code system to represent different kinds of design work within this design office, and also establish a set of procedures for filling out the input data, to make sure that each input data is in the right form. In the data processing side, we put formulas in the program to analyze and summarize the input data to become the output reports. In the meanwhile, we also combine the other files to produce different kinds of reports for different levels of management.

1.0 INTRODUCTION

In the past few years, due to the quick development among the Far East Asian countries, the manpower allocation and performance evaluation has been looked as an important phase in a company, especially in the design office where wants high technique level manpower.

This article deals with manpower problems faced by managers of design office in managing complex task forces in projects. The emphasis is on the implementation of a basic manpower control system which is essential for most project type design office. The system described here is designed to integrate all of the functions in manpower control and be used for computerized operation. It also can provide updated information to manpower management for decision making and directives. A system approach to this manpower control for functional and project managers is developed.

2.0 MANPOWER CONTROL SYSTEM

The manpower control system in this article has been designed to integrate the various functions of manpower management into a unified system to constantly moniter, forecast and report the manpower status of design offices and projects in relation to the existing manpower.

All of these functions are interrelated into a unified system as shown on the Functional Data Flow Chart, Figure 1.

The functions of manpower management can be separated into work code, manpower control, data collection, trend analysis, forecasting, reporting, and analysis. These seven major functions will be explained in following sections.

3.0 WORK CODE

The purpose of work code system is to establish a systemized work code structure. It provides a logical breakdown method for the elements of manpower resources in a project, so it is a positively identify system for these element.

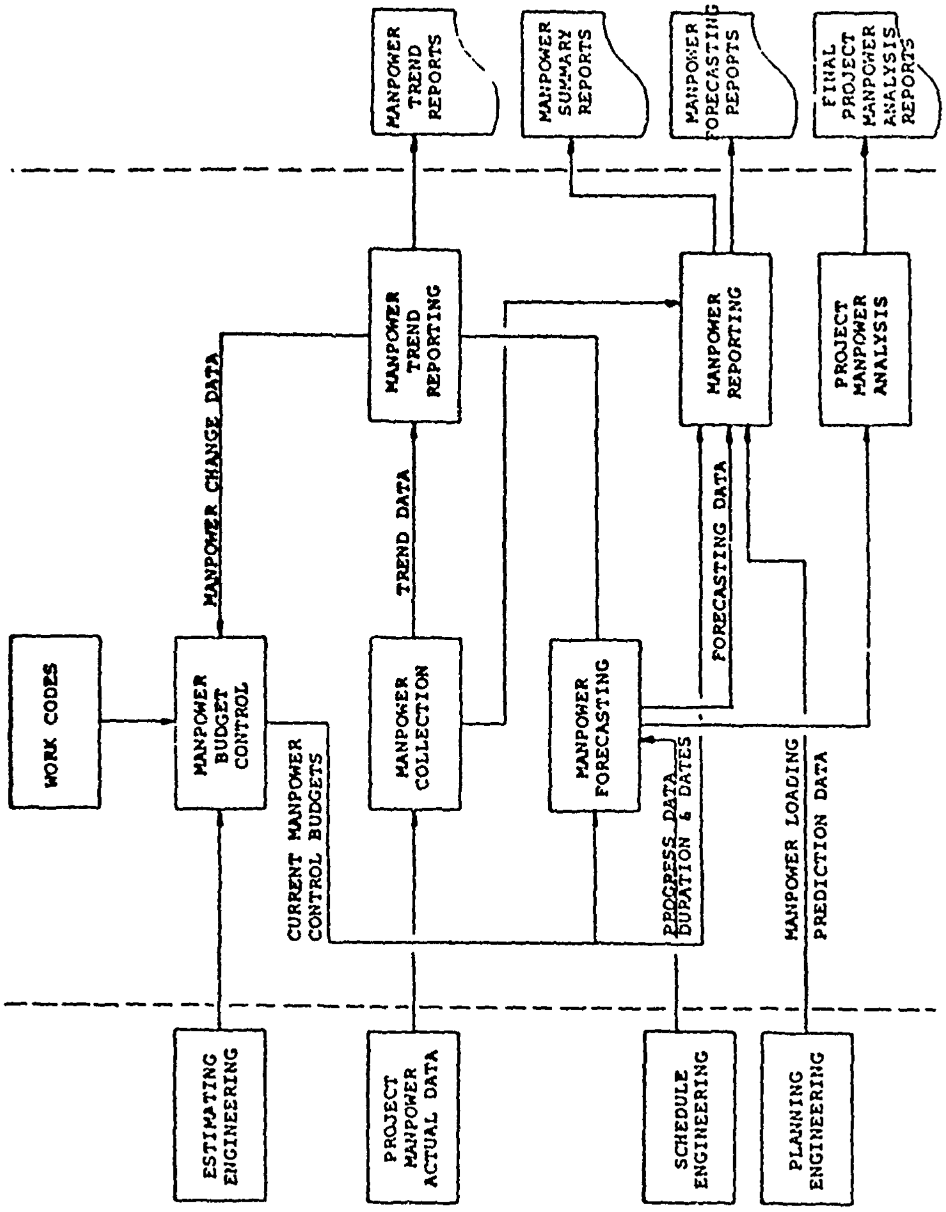


Fig. 1. Functional Data Flow Chart

The work code provides the means to collect project manpower data at the detail levels and moves this data a higher summary levels as required by the reporting functions. It can be modified as required to suit specific project reporting, project visibility requirements and data collection techniques.

4.0 MANPOWER CONTROL

The purpose of manpower estimate control is to properly initiate and maintain an up-to-date manpower control estimate for the project according to the work code. This is accomplished by closely monitoring and incorporating manpower change authorizations. Manpower control also provide a documented record for job changes.

The manpower control system executes under three different phases for each element in the work code structure:

. The initial control estimate
. Approved change orders
. The current control estimate

4.1 Initial Control Estimate

The initial control estimate is a detailed distribution of the job manpower according to the work plan and the work code breakdown. It should be made so detail as to permit direct comparisons between actual and estimate. The main basis of initial control estimate is the detailed estimate made in proposed estimate, and the historical manpower data also is an important reference.

4.2 Change Order

Changes orders are the only sources which will add or reduce the initial estimate. The change order should follow the formal procedure, and be approved by sponsor. Unless the written approval is made by project management and sponsor, no change is effected on initial estimate.

4.3 Current Control Estimate

The current control estimate always represents the initial control estimate, plus the value of all approved

change orders issued to date. The project personnel should record each change in job period to trace the estimate changing.

5.0 MANPOWER TRENDING

Manpower trending is the method used to give early visibility of manpower deviations during the operation of a project. So the trend is defined as any deviation from a planned project manpower or schedule which can affect the performance of the project.

The manpower trending report system provides the project management with timely information regarding manpower and schedule deviation from the approved project plan. This information can be used to the manpower planning decision and future manpower estimate.

The basic trend program should be consistent on all projects and follow the manpower trend report system. It is desirable that all responsible personnel within the projects initiate a trend report upon being aware of an actual or potential deviation to the program plan.

The manpower trending system consists trend analysis, trend control and trend report procedures, and is described as follow:

-Trend Analysis is an evaluation of the deviation that has been detected or will occur on a project. The analysis is done to determine the manpower and schedule effect of the deviation to date and project what the overall manpower and schedule effect of the manpower estimate will be at the end of the project.

-Trend Control procedure is the formal documentation of a trend, and will form the manpower trend history for a particular project file.

-Trend Report is the formal document for initiating the processing of a trend. It is used as an input document and also estimated the manpower increase or decrease caused by the deviation, projected through the completion of the project.

In the past few years, the computer application in design office has been developed so quickly that it already gives a large impact in design work. This trend should be trace carefully and to show what level it already influenced in design work.

6.0 DATA COLLECTION

The data collection system is the focal point for collecting all actual manpower as they occur on the project. It is the input data center for the manpower control system and provides direct interface with all other relevant functions in design office during project execution.

The purpose of the data collection system is to provide an efficient method to capture all manpower to date for engineering activities. The raw data is analyzed and assembled into different kinds of report and provides actual to date data for management.

The project actual data for computer input consist the cost center, employee's number and name, project number, work code, work class, and the hours charged by employee. In order to fill out the right input data, some procedures have been issued. Each procedure has stated as clear as possible for giving employee an easy way to fill out report. And each report shall be approved by project manager before send to computer center.

The schedule engineering will supply the estimated project schedule at the beginning of project, and will give the actual progress during the project is operated.

The final required input data for manpower forecasting is the manpower of high possibility proposal which come from planning engineering. This kind of proposal always be studied carefully by sales department.

7.0 MANPOWER FORECASTING

Manpower forecasting are divided into two parts, the first is project manpower prediction, the second is the manpower forecasting of total design office. These two

are netted together to form an effective manpower forecasting.

7.1 Manpower Prediction In Project

From the collected data of last section, we can analyze the project progress to date, the manpower to achieve this progress, and an evaluation of the future progress required to complete the project, and furthermore to predict the anticipated final project manpower.

The purpose of the project manpower prediction system is to constantly analyze manpower trends, actual manpower accruals, project progress and provide an accurate prediction for the manpower work to be done to complete the project.

7.2 Manpower Forecasting of Total Design Office

Combined manpower prediction in projects, manpower needed for high possible proposals and the existing manpower, many simulation methods can be used to forecast the manpower needed in the future of design office. This is an essential manpower planning of design office, it will give the management level a good decision making information.

8.0 MANPOWER REPORTING

The manpower reporting is the result of the integrated manpower management system; it also is the product of data processing in computer. The system will provide pertinent manpower information to company management, project management and supervisory functions. The report system is a graphic distribution chart, it provides the top-management a concise impression of reports distribution, and shows what reports will sent to whom. In general, there are required three kinds of report, the management level report, project level report and supervisory level report.

8.1 Management Level Report

The management level report of design office include the following and each report will be submitted monthly:

- Manpower prediction of design office for next two years
- Manpower utility status report
- Project manpower summary report

8.1.1 Manpower Prediction of Design Office for Next Two Years

This report is combined the graphic and numeric together, to show the available manpower and the predicted manpower load in future two years. It is a necessary data for design office manpower planning.

8.1.2 Manpower Utility Status Report

This report provides managers the manpower consumption of last month. It brief states the percentage of productive, R & D, estimating, and general manhours of each division, it also presents the overtime and leaving ratio of last month.

8.1.3 Project Manpower Summary Report

This report gives the general manager a brief status of each project, especial the project progress and prediction of manpower needed for completing each project.

8.2 Project Level Report

The monthly project progress and manhour consumption report presents the major work code summary of each project, it shows the progress of major group of work codes, its manhour consumption percentage, and predict to complete. The project manager can review their work from this report.

8.3 Control Level Report

The control level reports include the register logs of internal operation documents and the detail manhour consumption status. This control reports should be detailed to each work code and be classified the grade of manpower used in this code. It gives the department a guide for manpower allocation to each project.

9.0 PROJECT MANPOWER ANALYSIS

The purposes of project manpower analysis is to develop a historical manhour data from project execution into a unified system of data records which can be used as a base of proposal estimate, trend estimate for all future engineering design. In order to accomplish the above functions, a serious of analysis methods should be developed, and many kinds of result charts and tables have been established. The general consideration are MH per drawing sheet, MH per square feet of drawing sheet, MH per major equipment, and etc. The data needed for different kinds of estimating mehtod is the basic consideration of result charts and tables. Therefore, the data shall cover the detail, definite and module estimating method.

10.0 CONCLUSION

The system as stated in the above sections from the data flow chart to the description of each function is an outline of this system. In order to operate this system there are still have a lot of works to be done, especial the coding system for this system.

Because this manpower management system actual is subsystem of total project control system, and the project control system is also a subsystem of integrated corporate MIS system, so every coding system should be considered as a whole corporate. Therefore, when developing this system, one should consider the other important phases in this corporate, and then combine them together to become an integrated system in a corporate.

REFERENCES

1 Ludwig, E.E., APPLIED PROJECT MANAGEMENT FOR THE PROCESS INDUSTRIES, Gulf Publishing, Texas,1974.

2 Rase, H.F. and Barrow M.H., PROJECT ENGINEERING OF PROCESS PLANTS, John Wiley & Sons, London, 1958.

3 Guthrie, K.M. and Associates, MANAGING CAPITAL EXPENDITURES FOR CONSTRUCTION PROJECTS, Craftsman Book, California, 1977.

4 Martin, C.C., PROJECT MANAGEMENT: HOW TO MAKE IT WORK, AMA COM, New York, 1976.

5 Keating, C.J., LOOKING BACK AT COST ESTIMATING AND CONTROL, AACE Transaction, 1978.

6 Bauman, H.C., ENGINEERING COSTS, COST ENGINEER'S NOTEBOOK, West Virginia, 1978.

7 Patrascu, P., DETAILED CONTROL ESTIMATE, COST ENGINEER'S NOTEBOOK, West Virginia, 1978.

THE OPTIMAL INVENTORY SYSTEM OF A JUTE MILL

PAKORN ADULBHAN and SAEREE SVETASRENI

Industrial Engineering and Management Division
Asian Institute of Technology
Bangkok, Thailand

ABSTRACT. One of the most important factors concerning the management of many jute mills is the type of raw fibers used in manufacturing products. The primary reason is that between 60 and 70 percent of the total manufacturing cost of most finished products is attributed to the cost of raw fiber. This paper attempts to find the optimal inventory policy of the raw fibers for a jute mill so that the given production plan of the mill can be attained at the minimum inventory cost. From the given production plan, the economic combination of raw fiber types required in each week has been evaluated by a linear programming approach - then a dynamic programming model of a raw fiber inventory system is formulated so as to find the optimal inventory policy that provides the weekly raw fiber requirement to the mill at a minimum cost.

1. INTRODUCTION

Kenaf which is grown locally in the Northeastern provinces of Thailand is the raw material for manufacturing gunny bags and other related products of the jute mill under study. The kenaf can be divided into three quality grades namely A, B and C, with A being the highest quality. From previous records, large price fluctuations are typical for the jute trade, and the prices of the jute have been forecast [4] and shown in Fig. 1. The raw fibers fed into the mill will be graded again according to their quality into six grades, each of which is called "in-process material". They are R2, R3, SC, R4, R5 and RTC (ropes, tangles and cutting). From the previous records the expected compositions of raw fiber types are summarized in Table 1.

Table 1 The Compositions of Raw Fibers in % by Weight

	R2	R3	SC	R4	R5	RTC
A	42.56	35.20	14.04	6.20	-	2.00
B	-	37.78	14.80	30.72	14.70	2.00
C	-	-	-	56.07	40.65	3.28

These in-process materials are combined together in some fixed combinations (see Table 2), while passing through the production processes, to become one of the four in-process products. These in-process products are heavy yarn, light yarn, hessian and twine. Finally, these in-process products are again combined with each other in other fixed combinations (see Table 3) to produce the finished products. These finished products are heavy cee, hessian, twist and twine.

Table 2 The Composition of In-process Products in % by Weight

	R2	R3	SC	R4	R5	RTC
Heavy Yarn	-	-	10	50	20	20
Light Yarn	-	80	20	-	-	-
Hessian	90	10	-	-	-	-
Twine	70	30	-	-	-	-

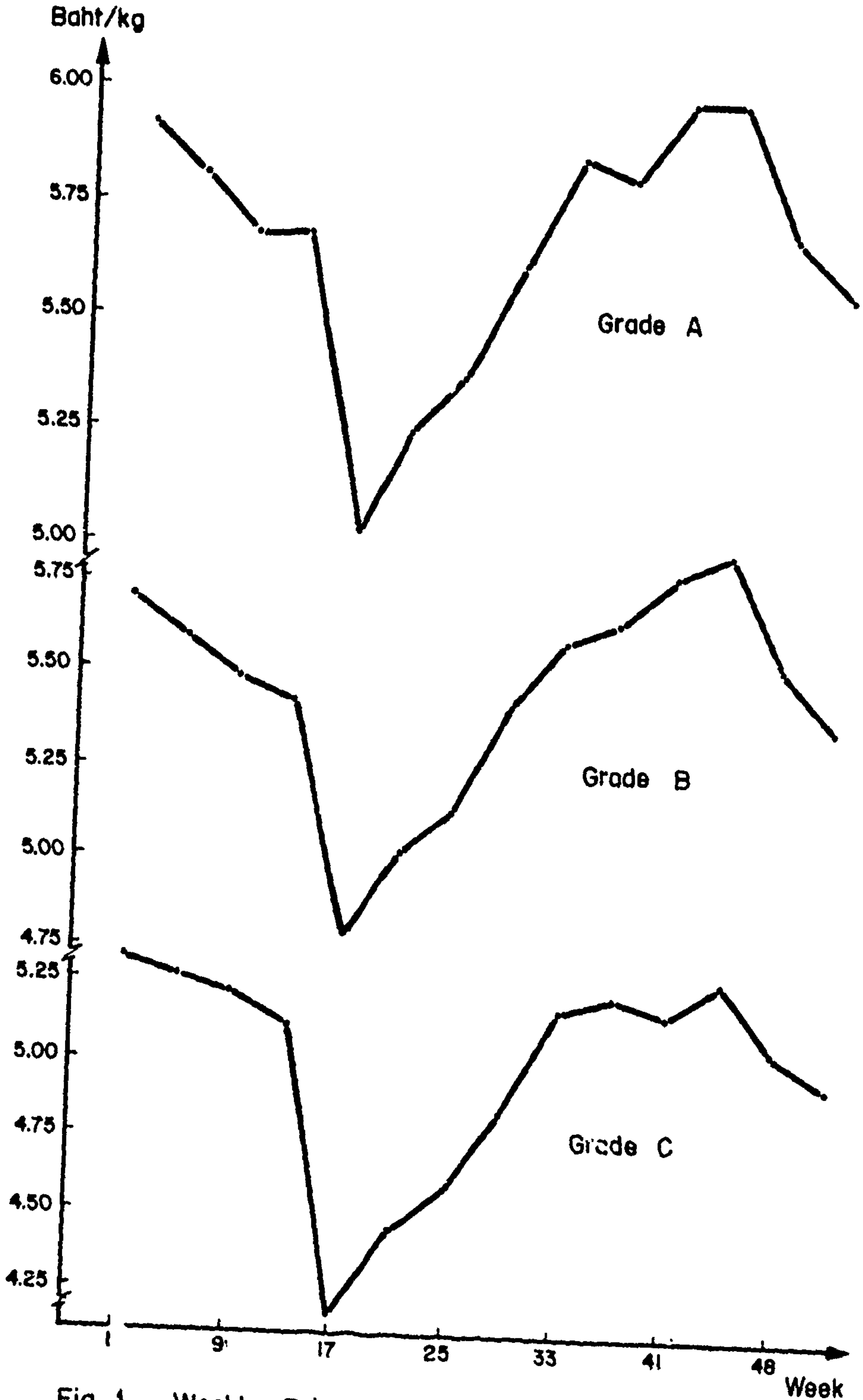


Fig. 1 Weekly Price of Raw Fiber Types in Year 1978

From the given production plan of the finished products (see Fig. 2), the required in-process products to be produced in each week so as to cope with the plan are computed by using Table 3.

Table 3 The Composition of Finished Products in % by Weight

	HEAVY YARN	LIGHT YARN	HESSIAN	TWINE
Heavy Cee	55	45	-	-
Hessian	-	-	100	-
Twist	10	90	-	-
Twine	-	-	-	100

The weekly requirements of in-process materials are then computed by help of Table 2.

Finally, from the weekly requirements for in-process materials, the optimal combination of raw fibers to be used in each week can be obtained from a linear programming (LP) model which will be formulated later. Before going into the details in formulating this model, it is quite helpful to discuss briefly the concept of linear programming.

Linear programming is a quantitative technique usually used in solving certain classes of allocation decision problems. Dantzig published his first paper on the subject in 1947. Since that time progress in the field has been rapid. The first applications were military in nature, but it was not long before it became apparent that there were important industrial applications as well.

Essential for a linear programming model is the requirement that the expression showing each of the restrictions should be linear with regard to each of the activities in question. Similarly, the expression that shows the relation of the activities to the objective function must be linear with respect to the activities. The iterative procedure known as the "Simplex Method" is normally used to obtain the solution to a linear programming problem, for it saves computational effort and is well suited for use in digital computers.

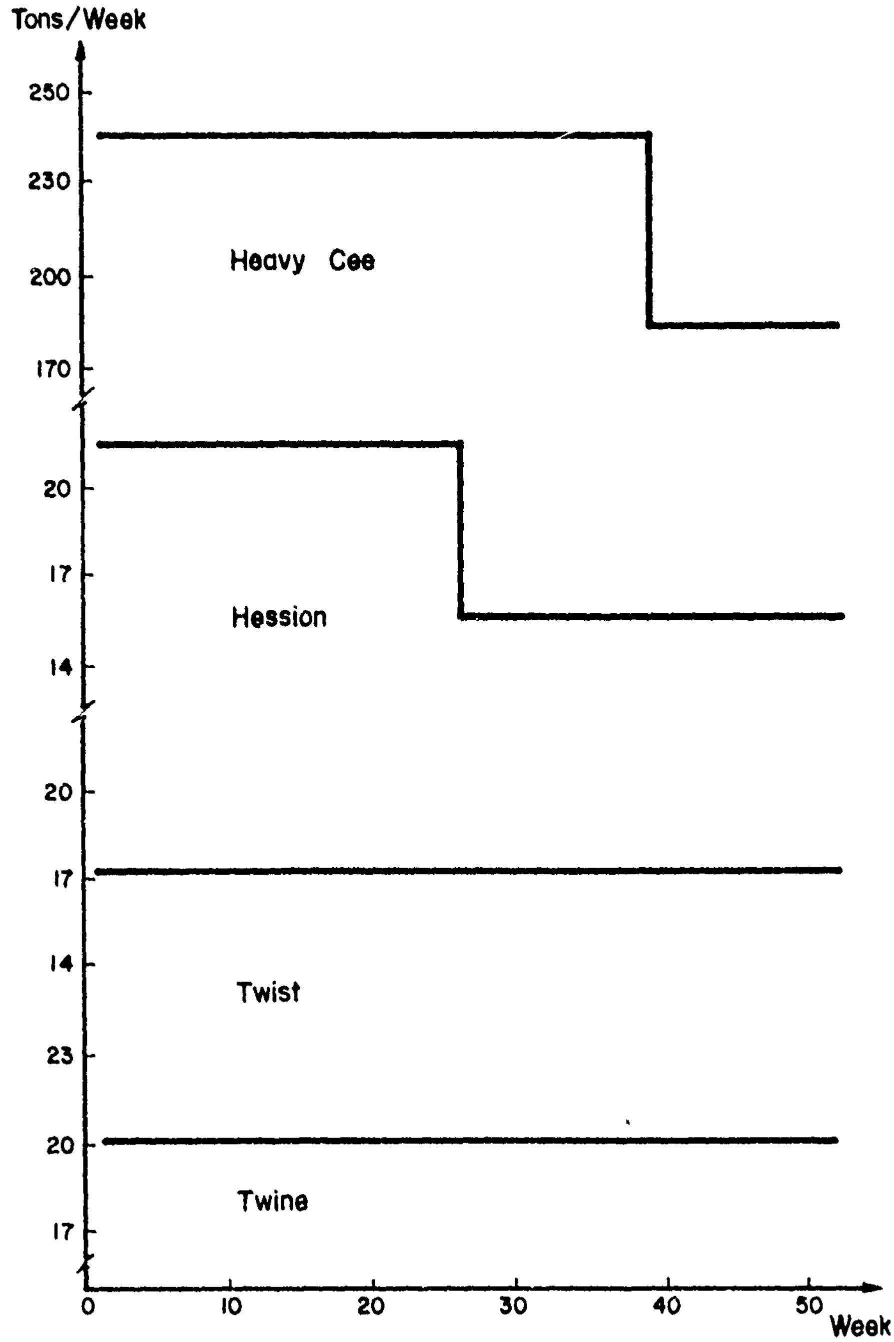


Fig. 2 Optimal Production Plan of Finished Products

The mathematical statement of a general form of linear programming is the following. Find the decision variables X_j which optimize the linear objective function,

$$Z = \sum_{j=1}^{n} c_j X_j \qquad (1)$$

subject to a set of restrictions which are referred to as constraints or restraints,

$$\sum_{j=1}^{n} a_{ij} X_j \quad (\leq, =, \geq) \quad b_i; \; i=1, 2, \ldots, m \qquad (2)$$

and $\quad X_j \geq 0 \quad$ for all $\quad j=1, 2, \ldots, n$

where in each constraint of (2) only one of three symbols $\leq$, $=$, $\geq$ appears. The coefficients a_{ij}, b_i and c_j are given as constants.

Once the linear programming model has been developed, what is referred to as the simplex method is used in determining the optimal (minimum or maximum) value of the objective function subject to the constraints involved.

2. FORMULATION OF LP MODEL

In any week, let Q_i = the quantity of raw fiber grade i, to be used in the mill, Tons, i = A, B or C

C_i = unit cost of raw fiber grade i, Baht/Ton

D_j = requirement of in-process material grade j, Tons, j = 1, 2,..., 6

P_{ij} = the percentage of in-process material grade j, in raw fiber grade i

The objective function of this model is to use the raw fibers so as to meet the given requirements for in-process materials in the corresponding week at the minimum cost,

i.e., Objective function:

$$\text{Minimize} \quad Z = \sum_i C_i Q_i; \qquad i = A, B \text{ and } C$$

Subject to the constraints:

$$\sum_i Q_i P_{ij} \geq D_j$$

where, $i = A$, B and C; $j = 1, 2, \ldots, 6$

From the weekly requirement for in-process materials D_j, that is computed previously, the unit prices of raw fibers (shown in Fig. 1) in the corresponding week and also the composition of each raw fiber grade obtained from Table 1 are then used as C_i and P_{ij} respectively in the above LP model. The model is then solved by using the subroutine[1] called "Optimization" which is available at the Regional Computer Center in Asian Institute of Technology (A.I.T.). The most economical combination of raw fibers to be used in each week can be obtained from this model and is shown in Fig. 3.

3. FORMULATION OF AN INVENTORY SYSTEM FOR THE RAW FIBER

From the weekly raw fiber requirement to be used in the mill (see Fig. 3), there must be an optimal inventory system that will provide enough raw fiber to the mill at the minimum inventory cost. In order to find this optimal inventory system, a dynamic programming (DP) model will be developed as follows,

The assumptions of the model are:

1. From a given production plan, the weekly requirement of each raw fiber type can be computed, and is assumed to be a deterministic demand. It is also assumed to appear at the beginning of the week.

---

1 It is in the package called "Subroutine Library-Mathematics (SL-MATH)"

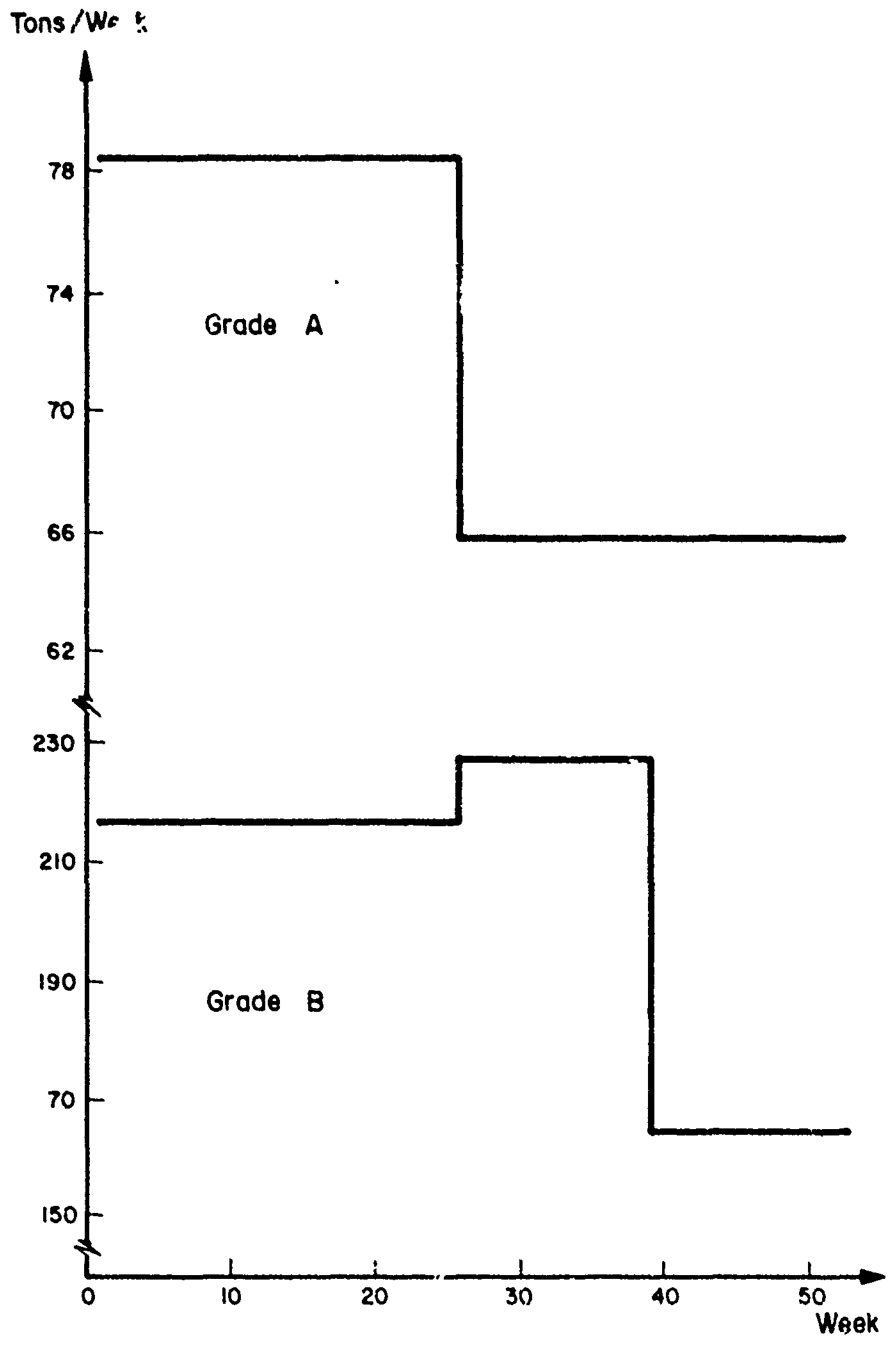


Fig. 3 Weekly Requirement of Raw Fibers

2. Orders are received instantaneously, and since the lead time in the procurement of raw fiber is relatively constant, it can be treated as a known and constant leadtime.

3. The cost of ordering per procurement is the same regardless of the order size.

'. Since the prices of raw fibers vary seasonally, in this model variable item cost will be considered. But the prices do not vary as a function of ordered quantity.

5. Only one type of raw fiber is considered at a time, in order to decrease the decision variables in the model.

6. Since the storage capacity of the mill is very high, it is assumed that the space as well as handling capacity and capital will not limit the amount of raw fiber bought and stored.[2]

7. At any time, for a given price of the raw fiber, there must be enough raw fiber to be purchased by the mill.

8. Since the raw fiber is the most important component in the jute mill industry, without it the mill has to stop operating, which causes a very big loss to the mill therefore in this model the shortage of raw fiber is not allowed.

9. At the beginning and at the end of the planning period, the inventory level of the raw fiber is zero.

The following symbols will be adopted in the model:

D_j - the requirement for raw fiber at the j^{th} week, Tons

C_j - unit cost of raw fiber at the j^{th} week, Baht/Ton

i - interest charged for capital investment in percent per week

C_o - ordering cost, Baht/Procurement

---

2 The constraint of storage capacity will be checked after getting the results.

C_h - holding cost, Baht/Ton/Week

L_i - inventory level of the raw fiber at the beginning of the i^{th} cycle, Tons

W_i - the week that the i^{th} cycle begins

Q_i - procurement quantity in the i^{th} cycle, Tons

NT_i - the number of weeks in the i^{th} cycle

TC_i - total inventory cost in the i^{th} cycle, Baht

CC_i - carrying cost in the i^{th} cycle, Baht

I_j - inventory level of the raw fiber at the j^{th} week, Tons

According to the assumptions of the model, the resulting inventory process may be represented graphically as in Fig.4, and the model can be formulated as follows:

$$PC_i = C_{W_i} \times Q_i$$

$$CC_i = C_h \times \bar{I}_i + \frac{C_{W_i} \times \bar{I}_i}{NT_i} \left\{ (1+i)^{NT_i} - 1 \right\} \times (1+i)^{53-(W_i+NT_i)}$$

where, $\bar{I}_i = \sum_{j=W_i}^{W_i+NT_i-1} I_j;$ for $I_j \geq 0$

In any i^{th} cycle, I_j can be computed from

$$I_{W_i+k} = L_i + Q_i - \sum_{j=W_i}^{W_i+k} D_j; \qquad k=0, 1, \ldots, (NT_i-1)$$

Then, $TC_i = C_o + PC_i + CC_i$

The objective of this model is to minimize the total inventory cost for the whole year (N cycles),

i.e., Minimize $TC = \sum_{i=1}^{N} TC_i$

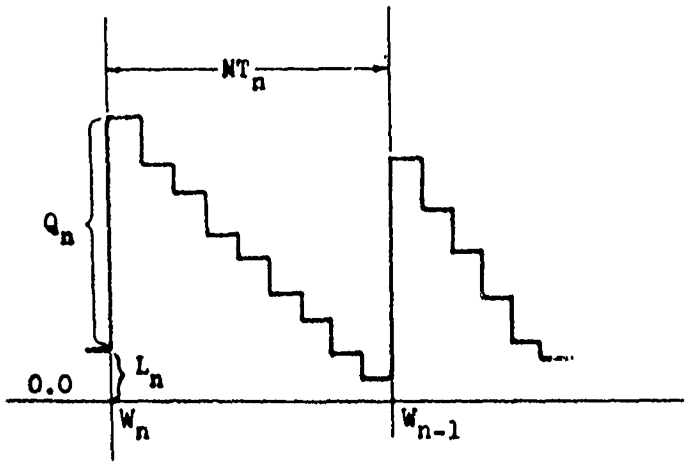


Fig. 4 Deterministic Inventory Process with Instantaneous Replenishment

Subject to the demand constraint,

$$\sum_{i=1}^{N} Q_i \geq \sum_{j=1}^{52} D_j$$

This problem can be formulated as a dynamic programming model as follows:

If each of the procurement actions is a stage, there will be N stages in the model starting from the last procurement action of the planning period (stage 1) to the first one (stage N).

Let W_n and L_n - the state variables of stage n

Q_n and NT_n - the decision variables of stage n

TC_n - the immediate return function of stage n

$f_n(W_n, L_n, Q_n, NT_n)$ - the cumulative return function of stage n

$f^*_{n-1}(W_{n-1}, L_{n-1})$ - the optimal cumulative return at stage n-1 when the state variables are W_{n-1} and L_{n-1}

The recursive relation of this model is

$$f_n(W_n,L_n,Q_n,NT_n) = TC_n + f^*_{n-1}(W_{n-1},L_{n-1})$$

and,
$$f^*_n(W_n,L_n) = \underset{Q_n,NT_n}{\text{Min.}}\ f_n(W_n,L_n,Q_n,NT_n)$$

where,

$$W_{n-1} = W_n + NT_n$$

$$L_{n-1} = L_n + Q_n - \sum_{j=W_n}^{W_n+NT_n-1} D_j$$

$$0 \leqslant Q_n \leqslant D; \qquad \text{(D = annual demand)}$$

$$0 \leqslant NT_n \leqslant 52$$

$$1 \leqslant W_n \leqslant 52$$

$$0 \leqslant L_n \leqslant D$$

Initial states at the first stage,

$$f^*_1(W_1,L_1) = \underset{Q_1,NT_1}{\text{Min.}}\ TC_1$$

where, $W_1 + NT_1 = 53$

and
$$L_0 = L_1 + Q_1 - \sum_{j=W_1}^{W_1+NT_1-1} D_j = 0$$

Initial states at the last stage,

$$W_N = 1, \qquad \text{and} \qquad L_N = 0$$

From the above model, a computer program using FORTRAN IV language is written so as to find the optimal solution of the model. The program listing and its flowchart are shown in the Appendix.

Inputs to the computer program for inventory model

The input parameters used in the computer program are listed below:

CO - ordering cost, Baht/procurement [4]

CH - holding cost, Baht/Ton/Week [4]

COST(J) - the price of raw fiber at the j^{th} week, Baht/Ton

DEMAND(J) - requirement of raw fiber at the j^{th} week, Tons

NF - the number of stages (procurement actions) desired in a year

Outputs of the computer program

The outputs obtained from this program are the optimal total annual inventory cost of raw fiber, and the corresponding procurement policy which will inform when the raw fiber should be purchased and in what quantity.

According to the optimal production plan (see Fig. 2), the economical requirements for raw fibers in this plan are computed (see Fig. 3) and fed into the inventory model, together with the weekly prices of raw fibers. The optimal procurement policy for the given production plan is computed and shown in Table 4.

Table 4 Optimal Procurement Policy for the Raw Fibers

	GRADE A		GRADE B	
	Procurement Quantity (Tons)	Week of Procurement	Procurement Quantity (Tons)	Week of Procurement
	561 561 2642	1 8 15	1328 2391 6957	1 7 18
Total	3764		10676	
Cost (Baht)	20,639,536		55,880,784	
Total Inventory Cost (Baht)			76,520,320	

4. CONCLUSION

The models developed in this paper have been shown to be useful when applied to the jute mill industry. The DP model, used in finding the optimal inventory system for raw fiber, considered the variation in the demands and also the price fluctuation of the raw fiber to be included in the model. It is not feasible in this paper to describe fully the estimation of the parameters used in the models. However, complete descriptions can be found in [4].

5. REFERENCES

[1] Buchan, J and Koenigsberg, E., SCIENTIFIC INVENTORY MANAGEMENT, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

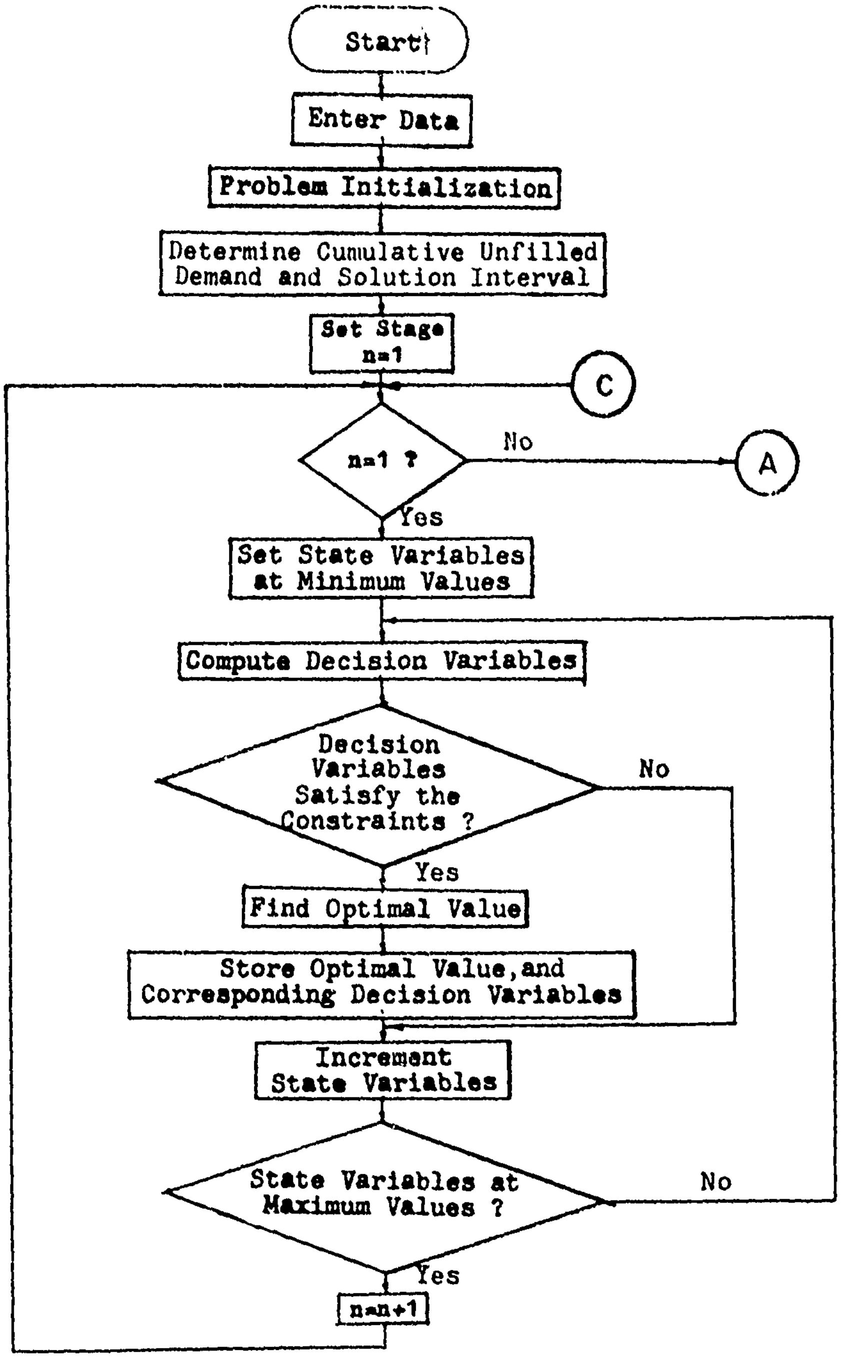
[2] Fabrycky, W.J. and Torgersen, P.E., OPERATION ECONOMY: INDUSTRIAL APPLICATIONS OF OPERATIONS RESEARCH, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1966.

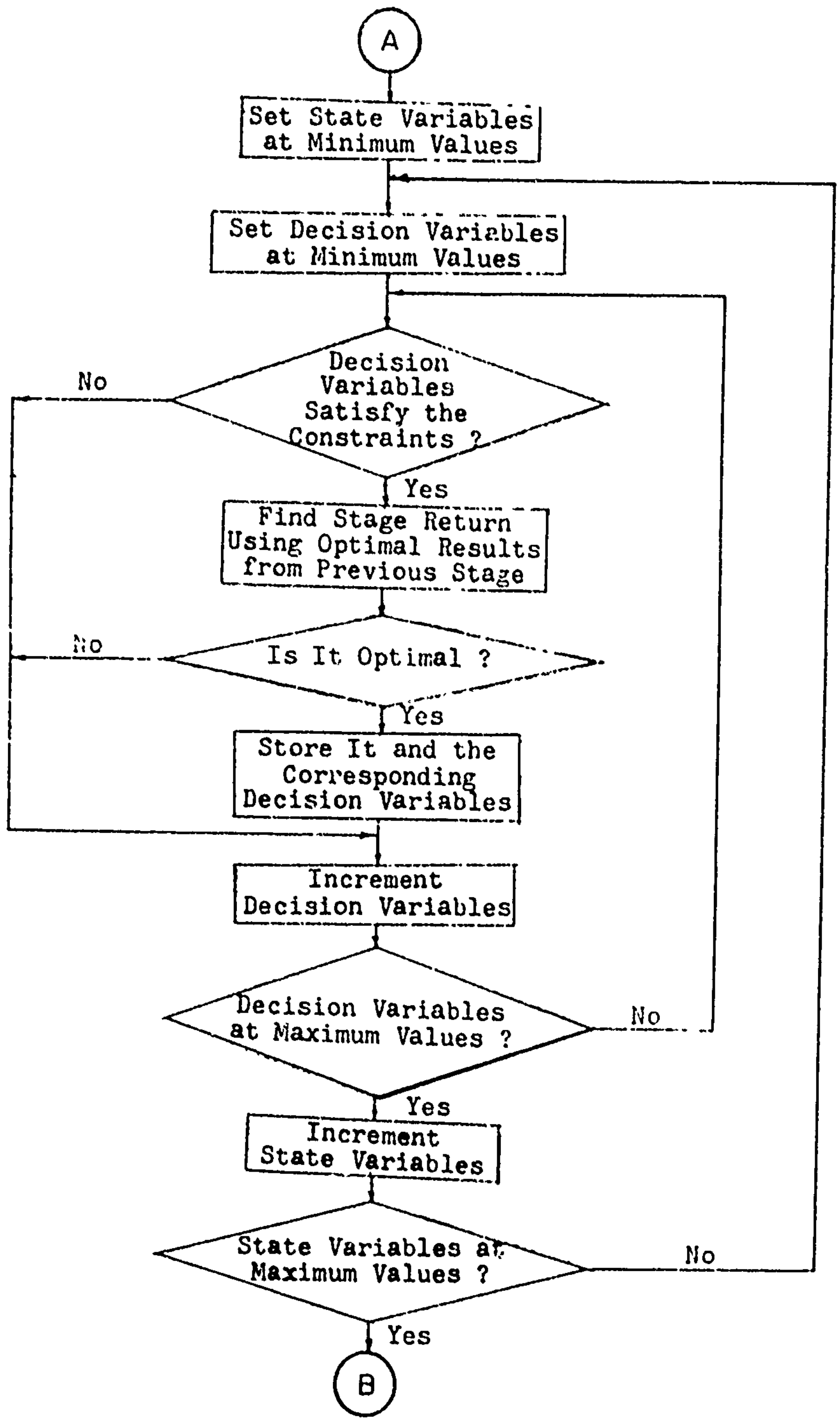
[3] Kuester, J.L. and Mize, J.H.; OPTIMIZATION TECHNIQUES WITH FORTRAN, McGraw-Hill Book Company, 1973.

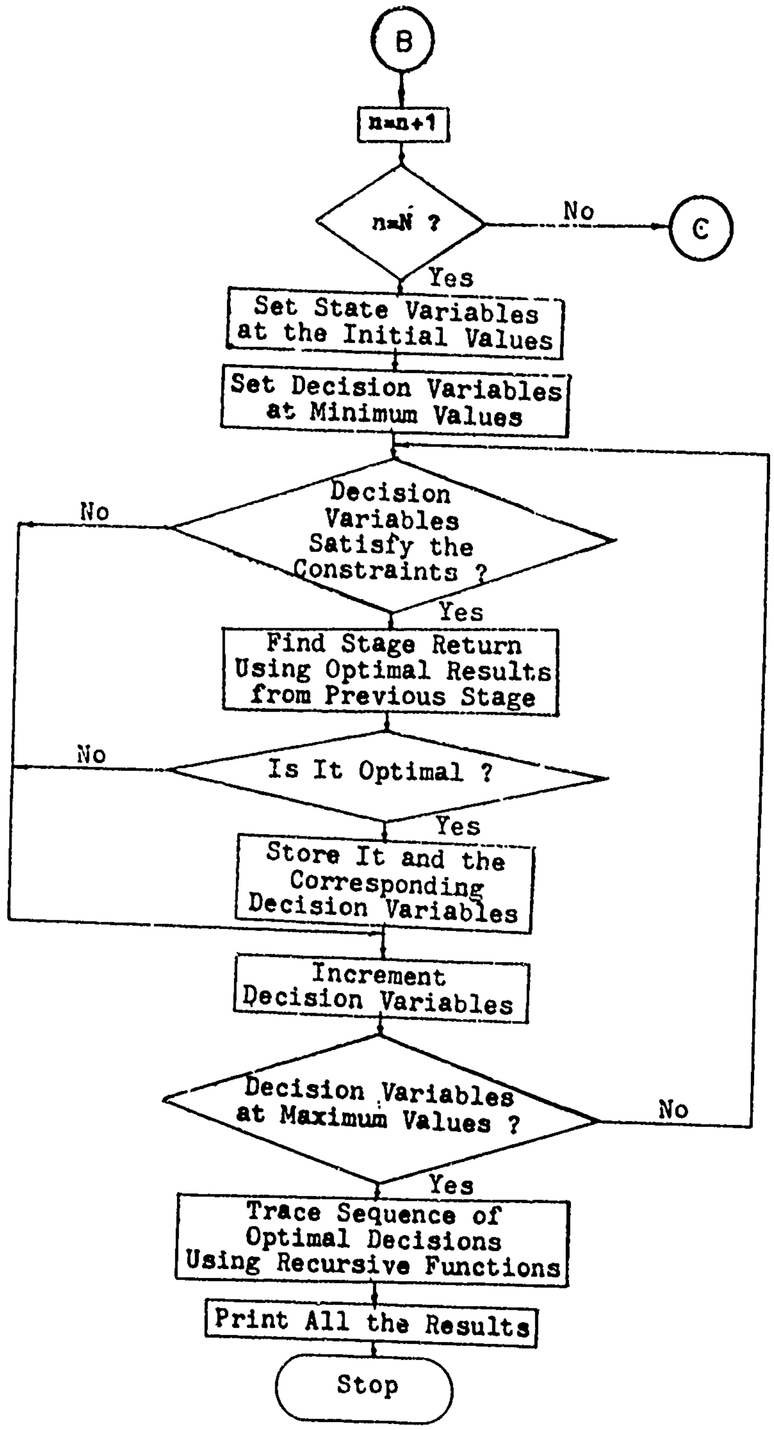
[4] Svetasreni, S., A PRODUCTION-SYSTEM STABILITY MODELLING FOR A JUTE MILL, Thesis, Asian Institute of Technology, Bangkok, Thailand, 1978.

APPENDIX

Computer Flow Chart for Optimal Inventory System of Raw Fibers




A
Set State Variables at Minimum Values
Set Decision Variables at Minimum Values
Decision Variables Satisfy the Constraints ?
No
Yes
Find Stage Return Using Optimal Results from Previous Stage
Is It Optimal ?
No
Yes
Store It and the Corresponding Decision Variables
Increment Decision Variables
Decision Variables at Maximum Values ?
No
Yes
Increment State Variables
State Variables at Maximum Values ?
No
Yes
B


B
n=n+1
n=N ?
No
C
Yes
Set State Variables at the Initial Values
Set Decision Variables at Minimum Values
Decision Variables Satisfy the Constraints ?
No
Yes
Find Stage Return Using Optimal Results from Previous Stage
Is It Optimal ?
No
Yes
Store It and the Corresponding Decision Variables
Increment Decision Variables
Decision Variables at Maximum Values ?
No
Yes
Trace Sequence of Optimal Decisions Using Recursive Functions
Print All the Results
Stop

COMPUTER PROGRAM FOR OPTIMAL INVENTORY SYSTEM OF RAW FIBER

```fortran
C
C.... INVENTORY SYSTEM OF RAW FIBER
C
      DIMENSION NAME(20),DEMAND(52),UNFDEM(52),FO(12,52,100),
     +          END(12,52,100),NTO(12,52,100),COST(52)
      COMMON DEMAND,CO,CH,COST
C
C.... PROBLEM INITIALIZATION
C
      DATA FO/62400*1000000000.0/
C
C.... ENTER DATA
C
      READ(5,101) NAME
      WRITE(6,201) NAME
      READ(5,102) DEMAND
      DO 1 I=1,4
         I2=13*I
         I1=I2-12
         WRITE(6,202) (DEMAND(K),K=I1,I2)
1     CONTINUE
      READ(5,102) COST
      DO 2 I=1,4
         I2=13*I
         I1=I2-12
         WRITE(6,202) (COST(K),K=I1,I2)
2     CONTINUE
      READ(5,*) CO,CH
      NF=3
      NF1=NF-1
C
C.... DETERMINE
C.... UNFDEM(NW) = UNFILLED DEMAND AT WEEK NW (TONS)
C....          C1 = SOLUTION INTERVAL(TONS)
C
      DX=0.0
      DO 3 N=1,52
         NW=53-N
         UNFDEM(NW)=DX+DEMAND(NW)
         DX=UNFDEM(NW)
3     CONTINUE
      DO 4 I=1,4
         I2=13*I
         I1=I2-12
         WRITE(6,202) (UNFDEM(K),K=I1,I2)
4     CONTINUE
      C1=UNFDEM(1)/100.0
      WRITE(6,203) CO,CH,C1
      DO 5 N=1,NF1
         N1=N-1
         DO 6 NW=1,52
```

```
               NLMAX=UNFDEM(NW)/C1+1
               NLDMAX=NLMAX+1
               IF(NLDMAX.GT.100) NLDMAX=100
               IF(N.NE.1) GO TO 7
C
C.....         FIND OPTIMAL VALUES AND DECISIONS
C.....         FOR EACH STATE OF THE FIRST STAGE
C
               DO 6 NLD=1,NLDMAX
                  NL=NLD-1
                  X1=C1*NL
                  QN=UNFDEM(NW)-X1
                  IF(QN.LT.0.0) QN=0.0
                  NT=55-NW
                  CALL STAGEN(NW,X1,QN,NT,GAIN)
                  FO(1,NW,NLD)=GAIN
                  QNO(1,NW,NLD)=QN
                  NTO(1,NW,NLD)=NT
6              CONTINUE
               GO TO 8
C
C.....         FIND OPTIMAL POLICY FOR EACH STATE
C.....         OF THE NEXT STAGE USING OPTIMAL
C.....         RESULTS FROM THE PREVIOUS STAGE
C
7              NTDM=55-NW
               DO 9 NLD=1,NLDMAX
                  NL=NLD-1
                  X1=C1*NL
                  QNMAX=UNFDEM(NW)-X1
                  NQMAXL=QNMAX/C1+2
                  DO 10 NQD=1,NQMAXL
                     NQ=NQD-1
                     QN=C1*NQ
                     DO 11 NTD=1,NTDM
                        NT=NTD-1
                        N1W=NW+NT
                        XN1L=X1+QN-(UNFDEM(NW)-UNFDEM(N1W))
                        IF(XN1L.LT.0.0) GO TO 10
                        N1L=XN1L/C1
                        N1LMAX=UNFDEM(N1W)/C1+1.0
                        IF(N1L.GT.N1LMAX) GO TO 9
                        N1LD=N1L+1
                        CALL STAGEN(NW,X1,QN,NT,GAIN)
                        F=GAIN+FO(N1,N1W,N1LD)
                        IF(F.GE.FO(N,NW,NLD)) GO TO 11
                        FO(N,NW,NLD)=F
                        QNO(N,NW,NLD)=QN
                        NTO(N,NW,NLD)=NT
11                   CONTINUE
10                CONTINUE
9              CONTINUE
8           CONTINUE
5        CONTINUE
```

```
C
C..... FIND OPTIMAL POLICY FOR THE INITIAL
C..... STATE OF THE LAST STAGE USING OPTIMAL
C..... RESULTS FROM THE PREVIOUS STAGE
C
      NW=1
      NL=0
      NLD=NL+1
      X1=0.0
      QNMAX=UNFDEM(NW)-X1
      NQMAXD=QNMAX/C1+2
      NTDM=53-NW
      DO 12 NQD=1,NQMAXD
         NQ=NQD-1
         QN=C1*NQ
         DO 13 NTD=1,NTDM
            NT=NTD-1
            N1W=NW+NT
            XN1L=X1+QN-(UNFDEM(1)-UNFDEM(N1W))
            IF(XN1L.LT.0.0) GO TO 12
            N1L=XN1L/C1
            N1LMAX=UNFDEM(N1W)/C1+1.0
            IF(N1L.GT.N1LMAX) GO TO 14
            N1LD=N1L+1
            CALL STAGEN(NW,X1,QN,NT,GAIN)
            F=GAIN+FO(NF1,N1W,N1LD)
            IF(F.GE.FO(NF,NW,NLD)) GO TO 13
            FO(NF,1,NLD)=F
            QNO(NF,1,NLD)=QN
            NTO(NF,1,NLD)=NT
13       CONTINUE
12    CONTINUE
C
C..... TRACE SEQUENCE OF OPTIMAL DECISIONS
C..... USING RECURSIVE FUNCTIONS,
C..... AND PRINT ALL THE RESULTS
C
14    NLD=1
      NW=1
      WRITE(6,204)
      DO 15 I=1,NF
         N=NF-I+1
         QN=QNO(N,NW,NLD)
         NT=NTO(N,NW,NLD)
         WRITE(6,205) N,FO(N,NW,NLD),QN,NT
         N1W=NW+NT
         NL=NLD-1
         X1=C1*NL
         XN1L=X1+QN-(UNFDEM(NW)-UNFDEM(N1W))
         N1L=XN1L/C1
         NLD=N1L+1
         NW=N1W
15    CONTINUE
      STOP
```

```
101   FORMAT(20A4)
102   FORMAT(8F10.0)
201   FORMAT(1H1,10X,20A4/)
202   FORMAT(1X,13(F6.2,2X))
203   FORMAT(/6X,'ORDERING COST',5X,'HOLDING COST',5X,'INTERVAL'/
     *        4X,F9.4,9X,E8.4,6X,F6.2//)
204   FORMAT(1H0,'THE OPTIMAL PROCUREMENT POLICY'/
     *        5X,'STAGE',5X,'CUMULATIVE INVENTORY COST',5X,
     *        'PROCUREMENT QUANTITY',5X,'USING PERIOD')
205   FORMAT(/6X,I2,13X,E13.2,10X,F9.2,15X,I2)
      END
C.... SUBROUTINE TO FIND GAIN=G(NW,X1,QN,NT)
      SUBROUTINE STAGEN(NW,X1,QN,NT,GAIN)
      DIMENSION DEMAND(52),COST(52)
      COMMON DEMAND,CO,CH,COST
      DATA AI/1.00209555/
C
C.... AI = A + I
C....  I = INTEREST CHARGED IN PERCENT PER WEEK
C
      FN(N)=AI**N
      IF(NT.NE.0) GO TO 1
      GAIN=0.0
      IF(QN.GE.1.0) GAIN=GAIN+CO+QN*COST(NW)
      RETURN
1     K=0
      J=NW
      XN=X1
      BARI=0.0
      XN=XN+QN
3     XN=XN-DEMAND(J)
      BARI=BARI+XN
      K=K+1
      IF(K.EQ.NT) GO TO 2
      J=J+1
      GO TO 3
2     NT1=53-NW-NT
      GAIN=QN*COST(NW)+BARI*COST(NW)*(FN(NT)-1)*FN(NT1)/NT
     *       +CH*BARI
      IF(QN.GE.1.0) GAIN=GAIN+CO
      RETURN
      END
```

APPLICATION OF MODELS IN ESTIMATING THE QUALITY COST FUNCTION.

1. A new division of quality costs.

The application of statistical models in controlling the quality costs of a product is not common at the moment. In my opinion, one of the main reasons for this is the traditional division of quality costs into the following components:

1. Prevention costs
2. Appraisal costs
3. Internal failure costs
4. External

From a model point of view, this division is not useful, because there is no clear relation between "the management controllable costs" (= 1. and 2.) and the result of the quality control, i.e. "The failure Costs". In this paper I will use the following division:

1. Prevention costs
2. Appraisal/inspection costs
3. Direct failure costs
4. Indirect failure costs
   a. Internal failure costs
   b. External failure costs

The direct failure costs are the failure costs incurred at the moment when a failure occurs.

The indirect failure costs are the extra failure

costs incurred because the failure was not discovered in the process where it happened. An example will clarify this division: A factory produces color-T.V.-sets. If a defective picture tube is manufactured the direct failure costs are rework or scrapcosts after the process which produced the defective part. If the defective part is not discovered after the relevant process extra failure costs are usually incurred. If the defective part is discovered only after assembly of the entire T.V.-set it is usually more expensive to rework the defective part (if rework is possible), than it was after the process which produced the defective part. If rework is not possible, it is necessary to scrap the defective picture tube, and it is evident, that scrapcosts are higher, if the defective picture tube has undergone one or more processes after the process which produced the defective part. All these extra failure costs are indirect internal failure costs. If the defect is not discovered in the factory, the customers are certain to discover it, and this means extra failure costs, such as transportation costs, "loss of goodwill" and "loss of reputation". These extra failure costs are the indirect external failure costs (Harrington, 1976, uses the term indirect costs as the "invisible costs", i.e. "Loss of Goodwill" and "Loss of Reputation"). A consequence of the division proposed in this paper is that there is a clear relation between management controllable costs and failure costs. If the appraisal costs are increased, the indirect failure costs will decrease. If the prevention

costs are increased a decrease in the direct failure costs may be expected. Models of these relations will be shown in this paper.

2. Models of the relation between inspection-costs and indirect failure costs.

The well-known economic condition for inspection of any lot is shown in (1) below:

$$n\pi C_I \geq n\ I \rightarrow$$

$$\pi \geq \frac{I}{C_I} \qquad (1)$$

where

n = Sample size

π = The defect proportion in the lot

C_I = The indirect failure costs per defective unit

I = The unit inspection costs

According to (1), the most economic decision is to inspect the whole lot, if the defect proportion is greater than the proportion between the unit inspection costs and the indirect failure costs. If the defect proportion is less than the proportion between I and C_I, the most economic decision is "no inspection".

2.1. Costmodels used in optimizing inspection of lots in any productionprocess.

The economic condition in (1) is so simple that, apparently, there is no need for any models.

The problem is however that the defect proportion is a random variable, which has to be estimated before a decision can be made. It is necessary to make an inference about the defect proportion. According to sample information, one of the following hypotheses must be selected:

$$H_o : \pi \leq \pi_o \quad H_1 : \pi > \pi_o$$

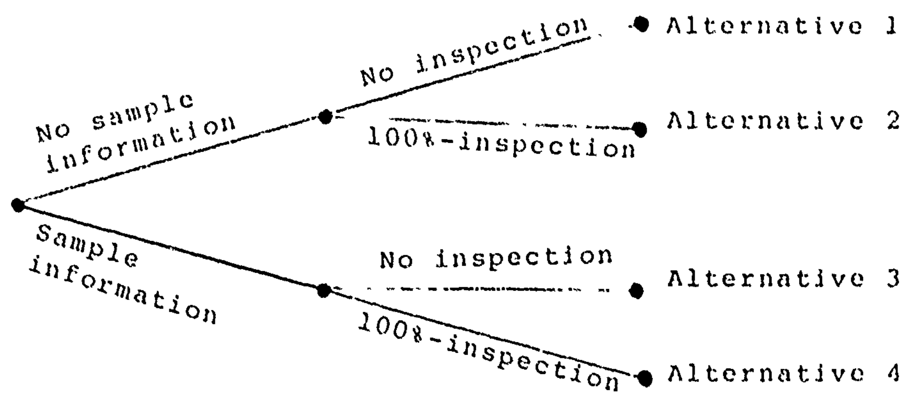
where

$$\pi_o = \frac{I}{C_I} \quad \text{(= the indifferent quality level)} \tag{2}$$

If H_o is accepted, the lot is also accepted, and there is no need for further inspection. If H_o is rejected, the lot is also rejected, and the whole lot has to be inspected.

If it is almost one hundred per cent certain, that H_o (or H_1) is true, inspection can be made without any sample information. So the decision problem for any lot is to choose one of the 4 decision alternatives in fig. 1.

Fig. 1: Decision alternatives for any lot.



If the random sample is a single sample plan, the criterion for rejecting H_0 is a sample containing more than c defective units. The sample size (=n) and c are found by minimizing the cost function in (3):

$$C(n,c)=\sum_{\pi}[nI+(N-n)\pi C_I\ Pr(X\leq c)+(N-n)I\ Pr(X>c)]\ f(\pi)$$

$$= nI+\sum_{\pi}[(N-n)\pi C_I \sum_{x=0}^{c} Pr(X=x)$$

$$+(N-n)I \sum_{x=c+1}^{n} Pr(X=x)]\ f(\pi) \qquad (3)$$

where:

N = The lot size

X = The number of defective units in the sample

$f(\pi)$ = The probability distribution of the defect proportion π

$$Pr(X=x) = \frac{\binom{N\pi}{x}\binom{N-N\pi}{n-x}}{\binom{N}{n}} \quad \text{(= the hypergeometric distribution)}$$

$1\leq n\leq N$

$0\leq c\leq n$

The minimizing of (3) is easily done by means of a computer.

Minimum of (3) should be compared with the <u>expected costs without sample information</u> (n=0):

$$\sum_{\pi} N\pi C_I\ f(\pi) = N\ C_I\ E(\pi) \qquad (4)$$

and the optimal solution is easily found.

The costfunction (3) is used when $f(\pi)$ is a discrete probability distribution. For convenience, the costfunction with $f(\pi)$ as a continuous probability distribution is shown in (5):

$$
\begin{aligned}
C(n,c) = \int_0^1 [nI + (N-n)\pi C_I \sum_{x=0}^{c} \Pr(X=x) \\
+ (N-n) I \sum_{x=c+1}^{n} \Pr(X=x)]\, f(\pi)\, d\pi \\
= nI + C_I (N-n) \sum_{x=0}^{c} \int_0^1 \pi \Pr(X=x) f(\pi)\, d\pi \\
+ (N-n) I \sum_{x=c+1}^{n} \int_0^1 \Pr(X=x) f(\pi)\, d\pi \qquad (5)
\end{aligned}
$$

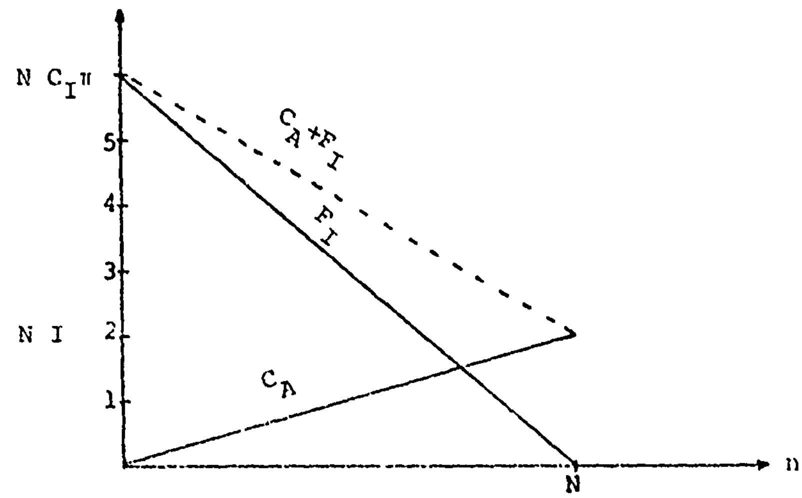
2.2. The relation between inspection costs and indirect failure costs, when inspecting lots for one quality-attribute (inspection in one productionprocess).

A product (or the components of a product) usually undergoes several productionprocesses, and in every process failure may happen. Some of the failures are of no or minor importance, and some are important. The direct and indirect failure costs are measures of the importance.

From (1) it follows that both the indirect failure costs and the inspection costs are linear functions. The relation between these costs for one quality-attribute is easily found, if the defect proportion is known. The construction of the relation is shown in figure 2.

Figure 2: Construction of the relation between the inspection costs and the indirect failure costs for one quality-attribute (process).

Figure 2a: The inspection and failure costfunctions



$C_A = n\ I$ = The inspection costs (appraisal costs)

$F_I = N\ C_I\ \pi(1-n/N)$ = The indirect failure costs

Figure 2b: F_I as function of C_A

2.3. The relation between inspection costs and indirect failure costs for lots of finished products (several productionprocesses)

Construction of the function showing the relation between inspection costs and indirect failure costs for lots of finished products is easy, if the following factors are known:

π_i = The defectproportion in lots from process no. i (i=1,2,3...L)

C_{Ii}= The indirect failure costs in process no. i (i=1,2,...,L)

I_i = The inspection costs per unit in process no. i (i=1,2,...,L)

N_i = The lotsize in process no. i (i=1,2,...,L)

The product-cost-function is built up from the costfunctions in the individual processes.
4 examples are shown in figures 3-6.

Because of the great number of processes which are usually necessary, when producing a product (for example a T.V.-set), it is useful to have a computer to construct the product-cost-function. If process-control-charts are used in the individual processes, the information about the defectproportions can be obtained from these controlcharts. If controlchart information is read by the computer, the construction of the costfunction can easily be automated.

Figures 3-6: Construction of the relation between the inspection costs and the indirect failure costs for a product with L attributes (processes):

Figure 3: Example 1

Figure 3a: The costfunctions of the individual processes (L=6 processes).

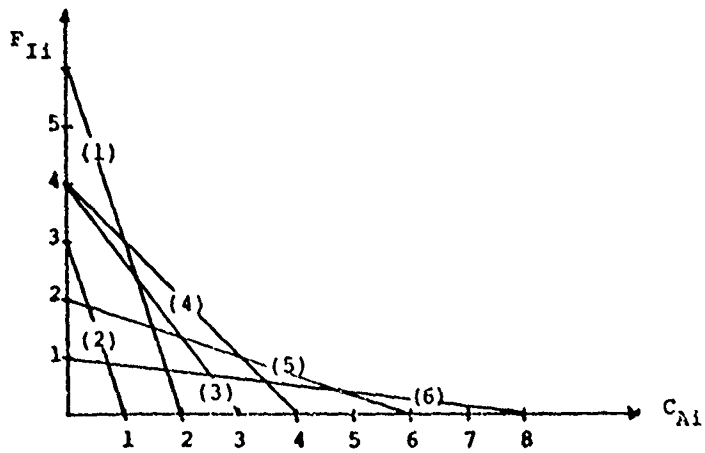


Figure 3b: The product-costfunction

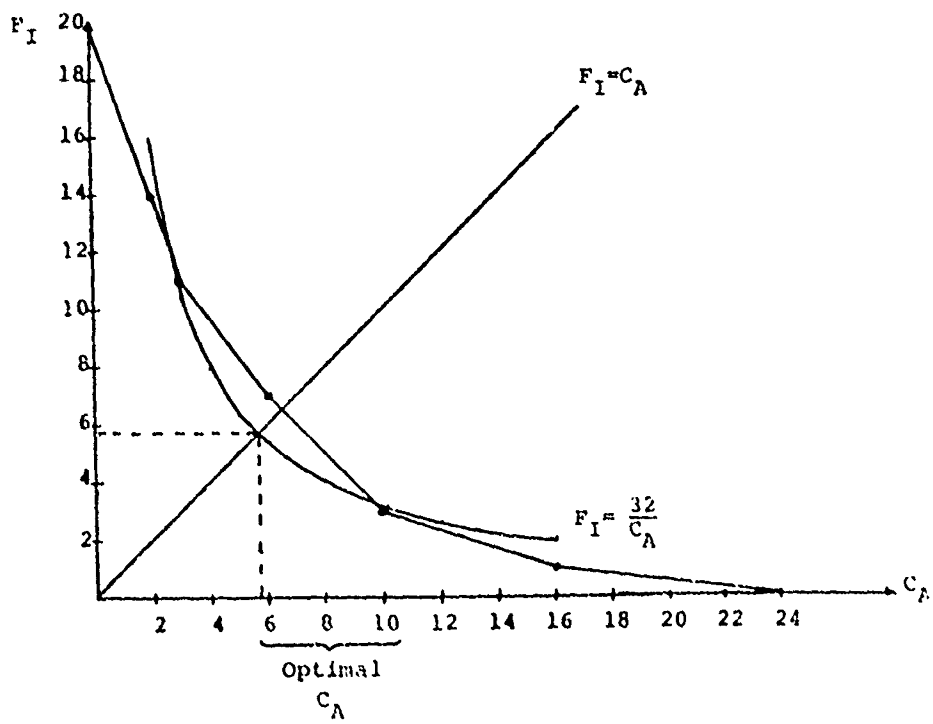


Figure 4: Example 2

Figure 4a: The costfunctions of the individual processes (L=5 processes)

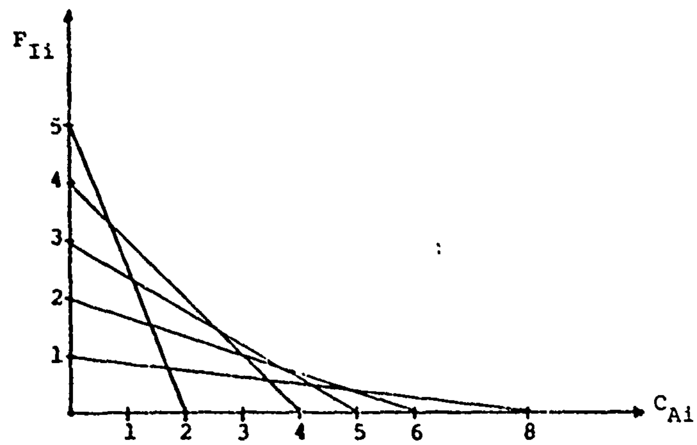


Figure 4b: The product-cost-function

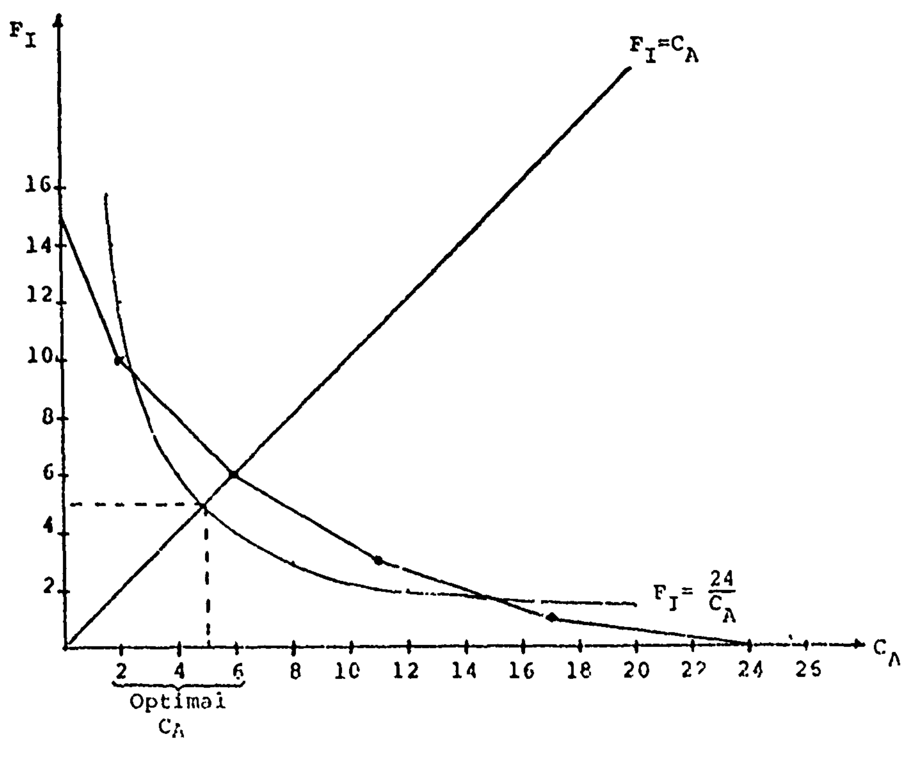


Figure 5: Example 3

Figure 5a: The costfunctions of the individual processes (L=5 processes)

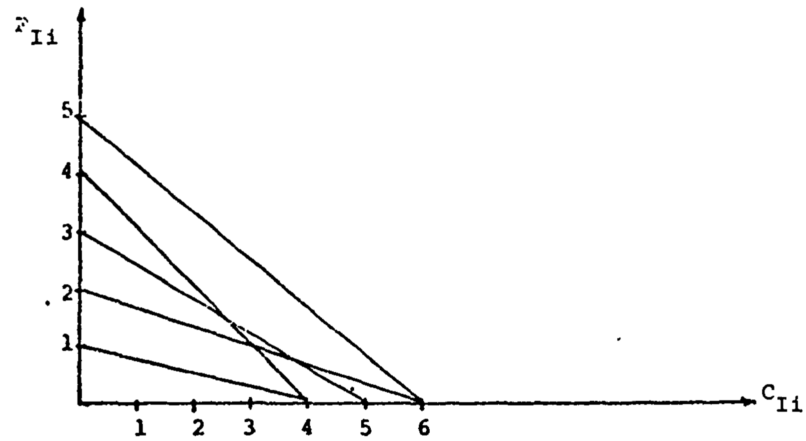


Figure 6b: The product-cost-function

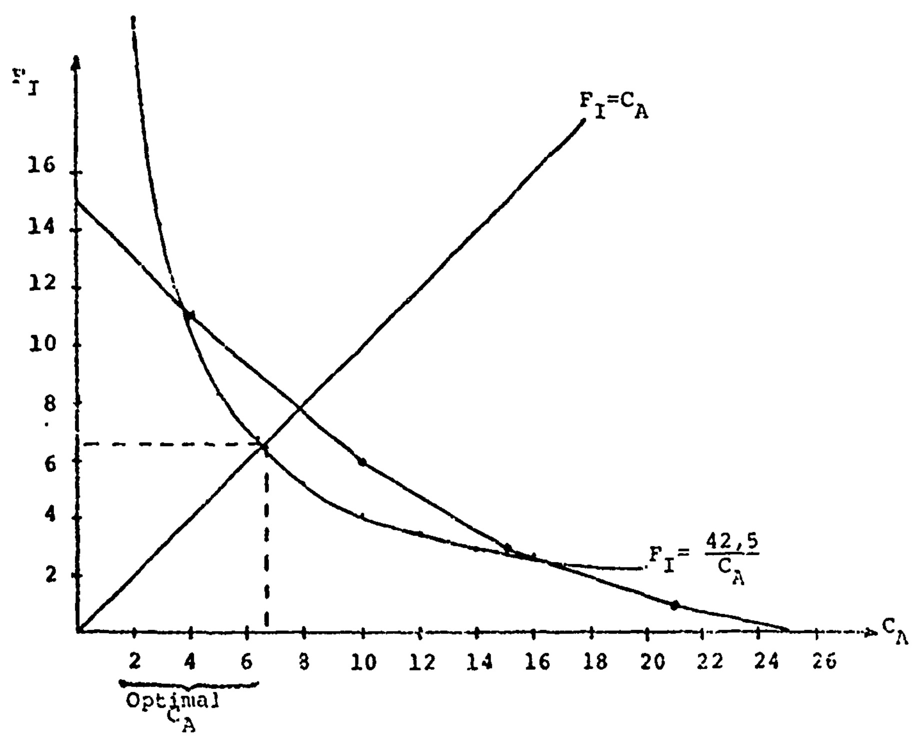


Figure 6: Example 4

Figure 6a: The costfunctions of the individual processes (L=5 processes).

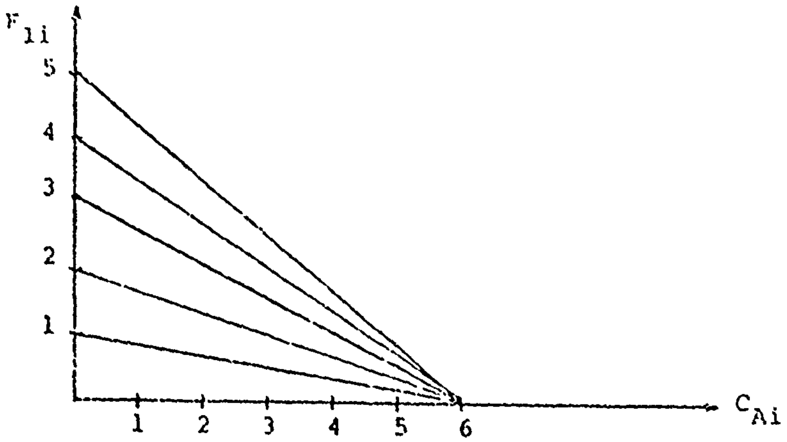
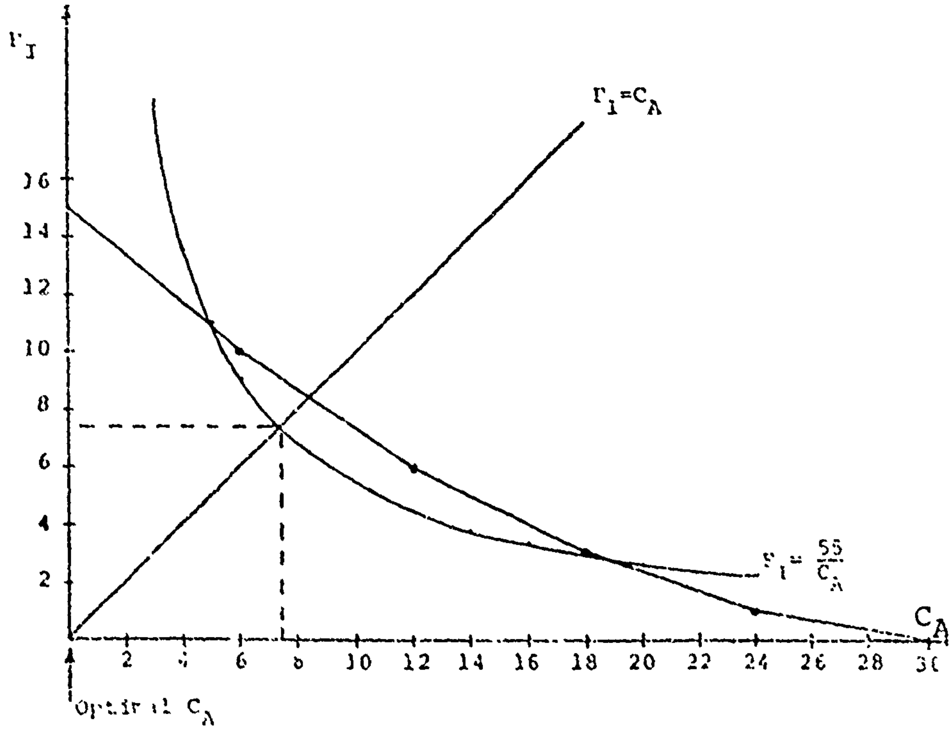


Figure 6b: The product-cost-function.



Compared with (1), the following deductions can be made:

1. If $\frac{dF_I}{dC_A} < -1$ It is economical to continue inspection

2. If $\frac{dF_I}{dC_A} > -1$ It is economical to stop inspection

3. If $\frac{dF_I}{dC_A} \doteq -1$ The inspection costs are equal to the expected indirect failure costs

If the defect proportions (π_i) are estimated from controlcharts or from other sorts of information, sample information from the lots is necessary when dF_I/dC_A is near -1.

Sometimes the hyperbola (Frey, 1976):

$$F_I = \frac{\Theta}{C_A} \tag{6}$$

where Θ is a positive constant

is proposed as a useful model, when the optimal inspection level has to be determined. In this model the optimal inspection level is the level, where $C_A = F_I$.

In fig. 3-6 Θ is estimated as [1] :

----

(1) This is a ML-estimate when the conditional probability distribution of F_I is:

$$f(F_I | C_A) = \frac{C_A}{\Theta} \quad e^{-\frac{C_A}{\Theta} F_I}$$

$$\hat{\theta} = \frac{\sum_i F_{Ii}\, C_{Ai}}{i} \tag{7}$$

where

(C_{Ai}, F_{Ii}) are the co-ordinates where the individual proces-cost-functions are connected ("Sample observations")

Only in fig. 4 is the optimal inspection level found by means of the hyperbola. In fig. 3 and fig. 5 the hyperbola-solution is near the optimal solution, but in fig. 6 the hyperbola-solution is far from optimum. From these examples it is clear that approximate models must be used discriminately.

2.4. An example where the indirect failure costs are expressed as a function of the productioncosts.

For certain products, the indirect failure costs are equal to the production costs in the processes between the process, where the failure happens, and the process where the failure is discovered. This is the case where failure makes it necessary to scrap the product. For products which are built up of many components, this assumption may be realistic for one ore more components. Perhaps this assumption is realistic for the main product as well, if it is too costly to rework the final product?

It is assumed in this example that if failure is

not discovered in the process where it happens, the failure will not be found before the finished products are tested at the earliest. If the failure is not discovered in the final test, a customer will find the failure, with extra failure costs as a result. The following notation will be used:

N_i = The lotsize produced in process no. i during the control periode (i=1,2,...,L)

π_i = The defect proportion in process no. i (i=1,2,...,L)

$q_i = n_i/N_i$ = The inspection-proportion in lots from process no. i (i=1,2,...,L)

π_{L+1} = The defectproportion of the finished products before the final inspection

q_{L+1} = The inspection-proportion when inspecting lots of finished products

a = The average extra charge per indirect failure money unit when a failure is discovered by a customer

P_i = The production-unit-costs in process no. i (i=1,2,...,L)

$$P = \sum_{i=1}^{L} P_i = \text{the production-unit-costs for the finished product}$$

$$V = \sum_{i=1}^{L} N_i P_i = \text{the value of the production in the controlperiod measured in productioncosts}$$

$$F_i = \sum_{j=1}^{i} P_j/P = \text{the degree of completion of the product after proces no. } i \qquad (i=1,2,\ldots,L)$$

With this notation the indirect failure costs of the control period (2) can be expressed as:

$$F_I = \sum_{i=1}^{L} N_i \pi_i (1-q_i) P (1-F_i) + N_L \pi_{L+1} (1-q_{L+1})(1+a) V \qquad (8)$$

If

$$N_1 = N_2 = \ldots N_L = N$$

the indirect failure costs can be expressed as:

$$F_I = V \left[\sum_{i=1}^{L} \pi_i (1-q_i)(1-F_i) + \pi_{L+1}(1-q_{L+1})(1+a) \right] \qquad (9)$$

----

(2) It is to be noted that failure may be discovered by the customers later than the controlperiod, so the external failure costs may "come to the surface" in a later period too.

The inspectioncosts of the controlperiod can be expressed as:

$$C_A = \sum_{i=1}^{L+1} N_i \, q_i \, I_i \tag{10}$$

where

$$N_{L+1} = N_L$$

I_{L+1} = inspection-unit-costs when inspecting the final product.

If

$$N_1 = N_2 = \dots = N_L = N_{L+1} = N$$

The inspectioncosts can be expressed as:

$$C_A = N \sum_{i=1}^{L+1} q_i I_i \tag{11}$$

The simple economic condition for inspection of any lot, which was shown in (1), can be expressed as:

$$\left.\begin{aligned} \pi_i &\geq \frac{I_i}{(1-F_i)P} \qquad (i=1,2,\dots,L) \\ \pi_{L+1} &\geq \frac{I_{L+1}}{(1+a)P} \end{aligned}\right\} \tag{12}$$

It follows from (8),(9),(10) and (11), that F_I and C_A are linear functions of the productionsize, if $\{q_i\}$ are held constant.

The fraction, S, is:

$$S = \frac{C_A}{V} \tag{13}$$

is a measure of inspection intensity. If S is raised, it may be expected that the growth in the indirect failure costs will be less than linear, when production increases.

If it is assumed that:

$$\sum_{i=1}^{L} \pi_i (1-q_i)(1-F_i)+\pi_{L+1}(1-q_{L+1})(1+a) = \frac{\Theta_A}{C_A/V} \tag{14}$$

where

Θ_A is a positive constant

(9) can be expressed as:

$$F_I = V \frac{\Theta_A}{C_A/V} = V \frac{\Theta_A}{S} \tag{15}$$
$$= V^2 \frac{\Theta_A}{C_A}$$

which is the hyperbola model introduced in (6). (15) can also be expressed as:

$$\frac{F_I}{V} = \frac{\Theta_A}{S} \tag{16}$$

which is a model showing these indirect failure costs as a fraction of productionvalue.

The parameter in (15) and (16) can be expressed

$$\Theta_A = \frac{C_A}{V} \left[\sum_{i=1}^{L} \pi_i \, (1-q_i)(1-F_i) + \pi_{L+1}(1-q_{L+1})(1+a) \right] \qquad (17)$$

which may be useful when estimating Θ_I.

3. Models of the relation between the prevention costs and the direct failure costs

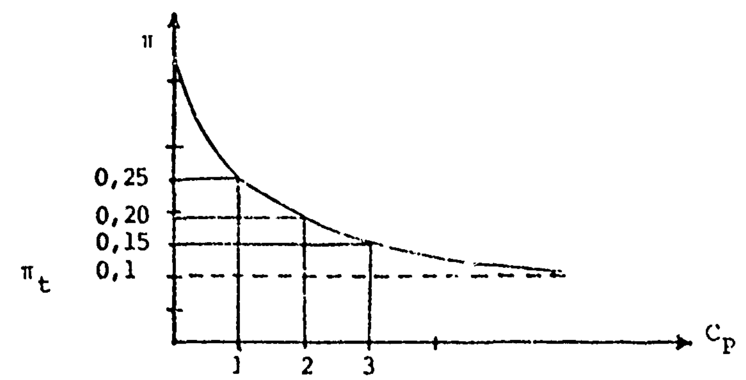
In section 2 of this paper it was shown how models could be used in controlling <u>the inspection costs</u> (= C_A) and <u>the indirect failure costs</u> (= F_I). These models must be used in connection with models of the relation between <u>the prevention costs</u> (= C_P) and <u>the direct failure costs</u> (= F_D), because the ultimate objective of quality control is to minimize <u>the total quality costs</u> (=C):

$$C = F_I + F_D + C_A + C_P \tag{18}$$

The minimizing of (18) is not easy neither in theory nor in practice. The problems are due to the fact that the defectproportions in any process are determined through the use of prevention costs, and as a consequence both the indirect failure costs and the inspection costs are dependent on the prevention costs. The principal relations are shown in fig. 7.

Figure 7: The relation between C_p, F_D, C_A and F_I in any productionprocess.

Figure 7a: The relation between π and C_p.

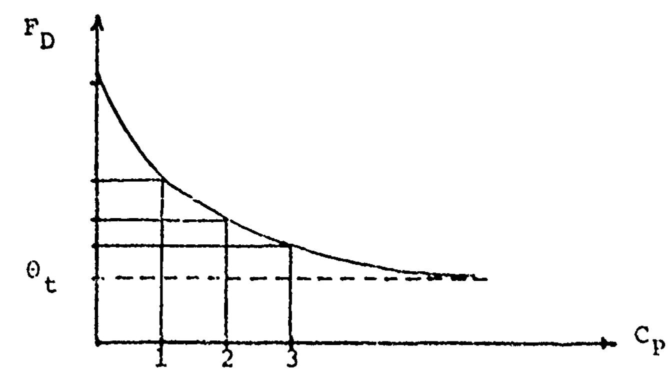


π_t = The lowest obtainable defectproportion with the technology used.

π = The defectproportion

C_P = The prevention costs

Figure 7b: The relation between F_D and C_p



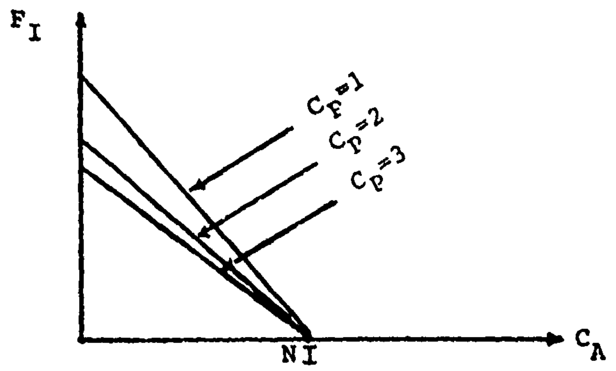
$F_D = N\pi C_D$ = The direct failure costs

N = Production size in units

C_D = The direct failure costs per defective unit

$\Theta_t = N\pi_t C_D$ = The lowest obtainable direct failure costs with the technology used

Figure 7c: The relation between F_I, C_A and C_P



F_I = (N-n) C_I = The indirect failure costs

C_A = The inspection costs

n = The inspection size

I = The unit inspection costs

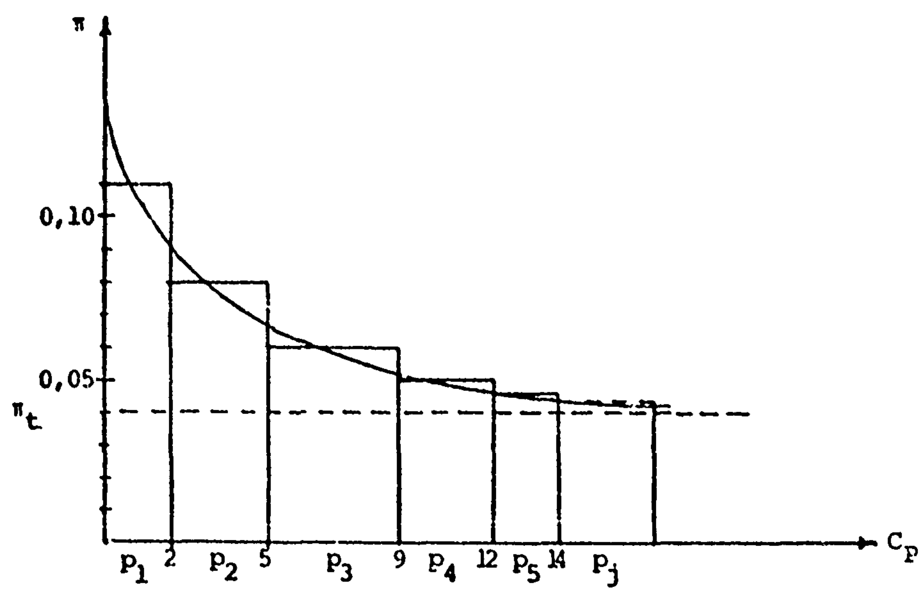
It is assumed that the defectproportion is 0.25, which is the result of prevention activities at $C_P=1$. When C_P is increased from "1" to "2", the defectproportion decreases to about 0.18. Because of the linear relationship between the defectproportion and the direct failure costs, the latter will decrease, following the same pattern as the defectproportion, which is shown in fig. 7b. The consequences for the relation between the inspectioncosts and the indirect failure costs are shown in fig. 7c. The decrease in inspection costs follows because it will be less economical to inspect the lots.

Besides the principal relations shown in fig. 10, there is another and perhaps greater problem with prevention costs. The prevention costs are the result of a number of different preventionprograms, which all have the purpose to decrease the defect-

proportion or to prevent an increase in the defect-proportion. Examples of preventionprograms are:

1. Education
2. Training
3. Implementation of process controlcharts
4. Implementation of QC-Circles

Figure 8: The principal relation between alternative preventionprograms and the defect proportion in any production process.



P_j = preventionprogram no. j (j=1,2,3,...)

The preventionprograms cost a fixed amount to carry out as shown in fig. 8. Besides, the results are not known in advance, and when a preventionprogram is implemented (for example P_2) it is not known if there may be programs which are more economical to implement. The choice and implementa-

tion of prevention programs presupposes hard efforts to find the dominant causes of defects, where the possible causes are (Juran, 1968,p.11-12):

1. Men
2. Materials
3. Methods
4. Machines
5. Management

In view of all the difficulties involved, it is perhaps preferable to use simple equations as (1) instead of more complex models. When evaluating prevention programs an economic condition for implementing a prevention program, P_j, is:

$$N[(\Delta\pi|P_j)(C_D+C_I) + I\Delta q] \geq (\Delta C_P|P_j) \tag{19}$$

where:

$(\Delta\pi|P_j)$ = The decrease in the defectproportion caused by prevention program P_j.

Δq = The decrease in the inspection-proportion

$(\Delta C_P|P_j)$ = The increase in prevention costs caused by preventionprogram P_j.

(18) can be expressed in the simpler way:

$$(\Delta\pi|P_j) \geq \frac{(\Delta C_P|P_j)/N}{C_D + C_I} - I\Delta q \tag{20}$$

which shows how much the defectproportion has to decrease for the quality program, P_j, to be

economical.

3.1. Two examples where the direct failure costs are expressed as a function of the production-costs.

In the notation used in section 2.4. if this paper, the two examples use the following direct failure costs ($=C_{Di}$) when a defective unit is produced in process no. i:

Example 1: $C_{Di} = P_i \qquad (i=1,2,\ldots L)$

Example 2: $C_{Di} = F_i\ P \qquad (i=1,2,\ldots L)$

In example 1 the direct failure costs are equal to the production costs per unit in the "failure-process", and in example 2, the direct failure costs are equal to the accumulated production costs. In complex products, built up of several components, perhaps example 1 is realistic for some components, and example 2 is realistic for the other components.

It is further assumed that:

$$N_1 = N_2 = N_3 \ldots = N_L = N$$

3.1.1. Example 1

Given the above-mentioned assumptions, and the notation used in the previous sections, the direct failure costs of the control period can be expressed as:

$$F_D = \sum_{i=1}^{L} N_i P_i \; \pi_i \tag{21}$$

$$= N \sum_{i=1}^{L} P_i \pi_i$$

$$= N \; P \sum_{i=1}^{L} \frac{P_i}{P} \pi_i = V \sum_{i=1}^{L} \frac{P_i}{P} \pi_i \tag{22}$$

If there is a defect level ($=\pi_{ti}$) determined by the technology in one or more processes, it may be useful to reexpress (22) as

$$F_D = V \sum_{i=1}^{L} \frac{P_i}{P} \pi_{ti} + V \sum_{i=1}^{L} \frac{P_i}{P} (\pi_i - \pi_{ti}) \tag{23}$$

or:

$$F_D = V \; \Theta_t + V \sum_{i=1} \frac{P_i}{P} (\pi_i - \pi_{ti}) \tag{24}$$

where [3]:

$\Theta_t = \sum_{i=1}^{L} \frac{P_i}{P} \pi_{ti}$ = a parameter, determined by technology, which is equal to the lowest obtainable direct failure costs per production-unit-cost. (25)

If it is assumed, that:

$$\sum_{i=1}^{L} \frac{P_i}{P} (\pi_i - \pi_{ti}) = \frac{\Theta_p}{C_{P/V}} \tag{26}$$

-----

(3) Θ_t is different from the usage in fig. 7.

(22) can be reexpressed as

$$F_D = V\,\Theta_t + V\,\frac{\Theta_P}{C_P/V}$$

$$= V\,(\Theta_t + \frac{\Theta_P}{C_P/V} \qquad (27)$$

where:

$$\Theta_P = \frac{C_P}{V} \sum_{i=1}^{L} \frac{P_i}{P} (\pi_i - \pi_{ti}) \qquad (28)$$

The relation between the prevention costs and the production value ($=C_P/V$) is a measure of the strength of the prevention activities [4]. If this measure is increased, it is assumed that one or more defectproportions will decrease in such a way that Θ_P will remain constant. It is further seen that (27) is a hyperbola, which can also be expressed as:

$$\frac{F_D}{V} = \Theta_t + \frac{\Theta_P}{C_P/V} \qquad (29)$$

The parameters of this model, as well as the parameter in (16), can be easily estimated and controlled by standard statistical methods. One of these methods has already been shown in (7).

------

(4) Perhaps C_P/V^x, where x is a positive parameter, is a better measure in some instances.

3.1.2. Example 2

Given the above assumptions, the direct failure costs can be expressed as:

$$F_D = \sum_{i=1}^{L} N_i F_i P \pi_i \tag{30}$$

$$= V \sum_{i=1}^{L} F_i \pi_i \tag{31}$$

$$= V \sum_{i=1}^{L} F_i \pi_{ti} + V \sum_{i=1}^{L} F_i (\pi_i - \pi_{ti})$$

or

$$F_D = V \Theta_t + V \sum_{i=1}^{L} F_i (\pi_i - \pi_{ti}) \tag{32}$$

where

$$\Theta_t = \sum_{i=1}^{L} F_i \pi_{ti} \tag{33}$$

If it is assumed that:

$$\Theta_p = \frac{C_p}{V} \sum_{i=1}^{L} F_i (\pi_i - \pi_{ti}) \tag{34}$$

(32) can be reexpressed as

$$F_D = V(\Theta_t + \frac{\Theta_p}{C_p/V}) \tag{35}$$

which is the same hyperbolamodel as (27). It must be recognized that the parameters of the models are not the same.

4. Optimization of the total quality costs

If the hyperbolamodels introduced in the previous sections of this paper are valid, (17) can be reexpressed as

$$C = V(\Theta_t + \frac{\Theta_P}{C_P/V}) + V\frac{\Theta_A}{C_A/P} + C_P + C_A \qquad (36)$$

Minimizing of (34) is complicated, because the two hyperbolamodels are dependent. The parameter Θ_A is dependent on the strength of the preventing activities $(= \frac{C_P}{V})$. When C_P/V is increased Θ_A will decrease. If the two models were independent it is easy to find the optimal solution through partial derivation:

$$\frac{\delta C}{\delta C_P} = \frac{\delta C}{\delta C_A} = 0 \quad \rightarrow$$

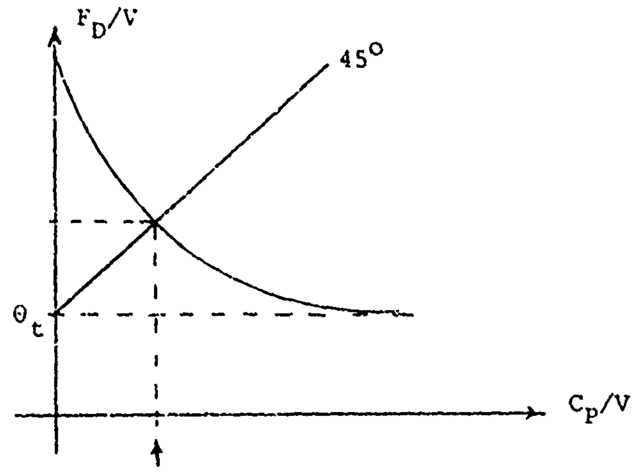
$$\text{"Optimal" } C_P = V\sqrt{\Theta_P} \qquad (37)$$

and

$$\text{"Optimal" } C_A = V\sqrt{\Theta_A} \qquad (38)$$

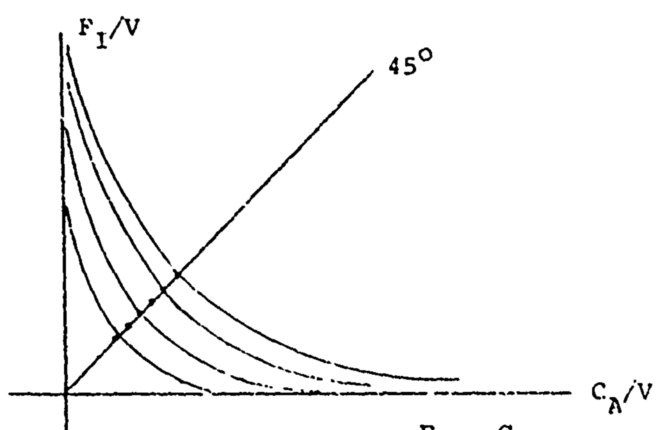
These "optimal" values are shown in figure 9 and figure 10. These values are practical start-values when optimazing C.

Figure 9: A principal draft of $F_D/V = \Theta_t + \frac{\Theta_P}{C_P/V}$



"Optimal"
C_P/V

Figure 10: Some principal drafts of $F_I/V = \frac{\Theta_A}{C_A/V}$



Optimal C_A/V is where $\frac{F_I}{V} = \frac{C_A}{V} = \sqrt{\Theta_A}$

When more preventionactivities are used, i.e. C_p/V is increased in relation to $\sqrt{\Theta_p}$, you "lose" money in fig. 9, but you "gain" money in fig. 10, because Θ_A is decreased. The economic condition for continuing the increase in C_p/V is:

$$2 \frac{d\sqrt{\Theta_A}}{d(C_p/V)} \leq \frac{d(F_D/V)}{d(C_p/V)} - d(C_p/V) \qquad (39)$$

There is no attempt in this paper to express a mathematical relation between Θ_A and C_p/V. In practice "trial-and-error-methods" may be used with (39) as a guideline.

5. Conclusion

From a new division of quality costs it has been shown how models of quality cost functions can be constructed. The examples shown may be or may not be realistic in practice. Further research is needed in constructing quality cost models with this new division. It is my hope that this new division of quality costs will give management a better possibility to control "Faults of the System" in an economic and scientific way (Deming, 1975, p. 3):

> "No improvement of the system, or any reduction of special causes of variation and trouble, will take place unless management attacks common causes with as much science and vigor as the production-workers and engineers attack special causes".

6. References:

1. W. Edwards Deming : "On Some Statistical Aids Toward Economic Production", Interfaces vol. 5, no. 4, aug. 1975.

2. Siegmar Frey : "Funktionelle Qualitätskostenbetrachtung", Technischer Rundschau, 10. august 1976.

3. H. James Harrington : "Quality Costs - The Whole and its Parts", Quality, may & june 1976.

4. J.M. Juran : "Quality Control Handbook", 2.nd.ed.,N.Y.1962 3.rd.ed., N.Y. 1974, McGraw-Hill.

GENERATING CORRELATED RANDOM VARIABLES

KIM ANDERSEN

Department of Operations Research
University of Aarhus
DK-8000 Aarhus C,
Denmark.

ABSTRACT. Working with stochastic models we usually make the assumption, that the stochastic variables are mutual independent. This assumption is made mainly because it is much difficult, if not impossible otherwise. Especially in many cases we even can't define the joint distribution for two stochastic variables with a given marginal distribution and with a given correlation between the variables. However, the assumption of independence is not always true, and a possible correlation between some of the stochastic variables may have a great influence on the results. This paper shows how to define a two dimensional stochastic variable, where the marginal distributions are uniform but with a given correlation between the variables. By use of this we obtain a very simple way to generate a sequence of random variables with a given marginal distribution and a given correlation between two consecutive variables.

List of Corrections

for

GENERATING CORRELATED RANDOM VARIABLES

by

KIM ANDERSEN

Read:	Instead of:	
$v_i + v_{i-1} < h$	$v_i + v_{i+1} < h$	
$0 < h \leq 1$	$0 \leq v_{i-1} \leq 1$	
$0 \leq v_{i-1} \leq 1$	$0 \leq v_i \leq 1$	
$0 \leq v_i \leq 1$		(2.2)

In formula (3.10) c is an integer, such that $0 < c < n$.

Read:	Instead of:	
$w_i + w_{i-1} \leq 2n - 2 - c$	$w_i + w_{i-1} \leq n - 1 - c$	
$2n - 2 - c < w_i + w_{i-1}$	$n - 1 - c < w_i + w_{i-1}$	
		(3.13)
$r + s = 2n - 1 - d$	$r + s = n - d$	
$r + s \leq 2n - 2 - d$	$r + s \leq n - 1 - d$	
$2n - 1 - d < r + s$	$n - d < r + s$	
		(3.34)

1. INTRODUCTION

Let us assume we want to generate a sequence of correlated random variables

$$x_1, x_2, \ldots, x_{i-1}, x_i, \ldots \tag{1.1}$$

with the joint density for x_{i-1} and x_i given as

$$f(x_i, x_{i-1}) = f(x_i | x_{i-1}) f(x_{i-1}) \tag{1.2}$$

and with a correlation between two consecutive variables equal to r_x.

From (1.2) we can find the conditional distribution

$$G(x_i) = F(x_i | x_{i-1}) = \frac{\int_{-\infty}^{x_i} f(x, x_{i-1})dx}{f(x_{i-1})} \tag{1.3}$$

and by use of the inverse function, G^{-1}, we can generate the sequence (1.1) by

$$x_i = G^{-1}(u_i, x_{i-1}) \tag{1.4}$$

where

$$u_1, u_2, \ldots, u_i, \ldots \tag{1.5}$$

is a sequence of uncorrelated uniform distributed random variables in the interval (0,1).

Usually, however, we only know the marginal density in (1.2) and r_x, and even if we did know the joint distribution in (1.2), it normally is extremely difficult, if not impossible, to find the inverse conditional distribution in (1.4). Given the marginal distribution, F, and r_x, in case we can generate a sequence of correlated uniform distributed random variables in the interval (0,1)

$$v_1, v_2, \ldots, v_i, \ldots \tag{1.6}$$

with a correlation between two consecutive variables equal to let us say r_v, then we can generate the sequence (1.1) by the usual inverse transformation

$$x_i = F^{-1}(v_i) \tag{1.7}$$

given that we know how r_x depends on r_v. If we can find the inverse function in (1.7), the only thing we need in this case is to generate the sequence in (1.6).

2. UNIFORM DISTRIBUTION WITH CORRELATION

Obviously, there exist many two dimensional distributions for v_i and v_{i-1} with the same uniform marginal distributions and with a correlation, r_v, between v_i and v_{i-1}. One of those is the distribution with the density

$$t(v_i,v_{i-1}) = t(v_i|v_{i-1})t(v_{i-1}) \tag{2.1}$$

where (see Fig. 2.1).

$$t(v_i|v_{i-1}) = \left\{\begin{array}{lll} \frac{1}{2h} & \text{if} & \left\{\begin{array}{l} h \le v_i + v_{i-1} \le 2-h \\ -h \le v_i - v_{i-1} \le h \end{array}\right. \\ \frac{1}{h} & \text{if} & v_i + v_{i+1} < h \\ \frac{1}{h} & \text{if} & 2-h < v_i + v_{i-1} \\ 0 & \text{else} & \end{array}\right\} \begin{array}{l} 0 \le v_{i-1} \le 1 \\ 0 \le v_i \le 1 \end{array} \tag{2.2}$$

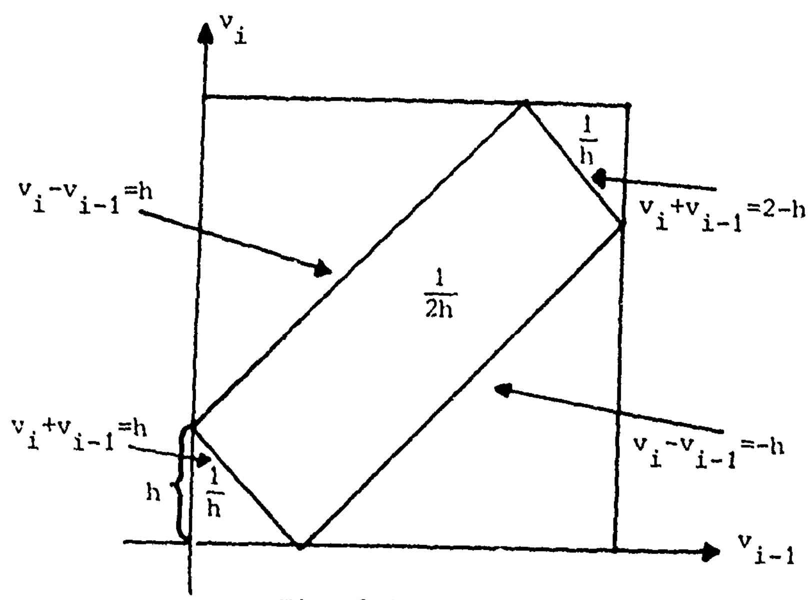


Fig. 2.1.

If the marginal density for v_{i-1} is uniform and given by

$$t(v_{i-1}) = 1 \qquad 0 \leq v_{i-1} \leq 1 \tag{2.3}$$

we can deduce that the marginal density for v_i is

$$\int_{\max(0,v_i-h)}^{\min(1,v_i+h)} t(v_i|v_{i-1})t(v_{i-1})dv_{i-1} = 1 \qquad 0 \leq v_i \leq 1 \tag{2.4}$$

That is to say, the marginal distribution for v_i is uniform, too.

As

$$Ev_iv_{i-1} = \int\int v_iv_{i-1}t(v_i,v_{i-1})dv_idv_{i-1} \tag{2.5}$$

$$= \frac{1}{2h}\int_0^1\int_0^1 v_iv_{i-1}dv_idv_{i-1} - \frac{2}{2h}\int_h^1\int_0^{v_{i-1}-h} v_iv_{i-1}dv_idv_{i-1}$$

$$+ \frac{1}{2h}\int_0^h\int_0^{h-v_{i-1}} v_iv_{i-1}dv_idv_{i-1}$$

$$+ \frac{1}{2h}\int_{1-h}^1\int_{2-v_{i-1}-h}^1 v_iv_{i-1}dv_idv_{i-1}$$

$$= \frac{1}{2h}\left[\frac{1}{4} - \frac{2}{24}(-h^4 + 6h^2 - 8h + 3) + \frac{h^4}{24}\right.$$

$$\left. + \frac{1}{24}(h^4 - 8h^3 + 12h^2)\right]$$

$$= \frac{1}{12}[h^3 - 2h^2 + 4]$$

and

$$Ev_i = Ev_{i-1} = \frac{1}{2}$$
$$\text{Var } v_i = \text{Var } v_{i-1} = \frac{1}{12} \tag{2.6}$$

we have for the covarians between v_{i-1} and v_i

$$\text{cov}(v_{i-1},v_i) = \frac{1}{12}[h^3 - 2h^2 + 1] \tag{2.7}$$

and hence for the correlation between v_{i-1} and v_i

$$r_v(h) = h^3 - 2h^2 + 1 = (1-h)(1+h-h^2) \tag{2.8}$$

From (2.8) we see that $r_v(h)$ is a decreasing function on h with $r_v(0) = 1$ and $r_v(1) = 0$.

To generate the sequence in (1.6) we can use the following algorithms

$$\begin{aligned}&v_i := \text{uniform}(v_{i-1} - h,\ v_{i-1} + h);\\ &\underline{\text{if}}\ v_i < 0\ \underline{\text{then}}\ v_i := -v_i\ \underline{\text{else}}\\ &\underline{\text{if}}\ v_i > 1\ \underline{\text{then}}\ v_i := 2 - v_i;\end{aligned} \tag{2.9}$$

where uniform (a,b) generates a uniform distributed random variable in the interval (a,b).

From the fact that

$$\text{cov}(1 - v_{i-1}, v_i) = -\text{cov}(v_{i-1}, v_i) \tag{2.10}$$

we can generate the sequence in (1.6) with a negative correlation by substituting $1 - v_{i-1}$ for v_{i-1} in (2.9).

3. GENERAL DISTRIBUTION

Given the marginal distribution, F, we can generate the sequence (1.1) by the transformation

$$x_i = F^{-1}(v_i) \tag{3.1}$$

where v_i is generated by the algorithms in (2.9).

To find how the correlation, r_x, between x_{i-1} and x_i depends on h, we have to find

$$\begin{aligned}Ex_ix_{i-1} &= \int\int x_ix_{i-1}\ f(x_i, x_{i-1})dx_idx_{i-1}\\ &= \int\int F^{-1}(v_i)F^{-1}(v_{i-1})t(v_i, v_{i-1})dv_idv_{i-1}\end{aligned} \tag{3.2}$$

and then use

$$r_x = \frac{Ex_ix_{i-1} - (Ex_i)^2}{\text{Var}\ x_i} \tag{3.3}$$

However, it is usually very difficult to compute both (3.1) and (3.2), even by means of numerical methods. Hence, it is appropriate to make a small change in the above mentioned method and obtain the quantile transformation.

3.1. The Quantile Transformation

Instead of using the inverse transformation every time we want to generate a random variable with the distribution, F, we can for n different values of p_j, for instance

$$p_j = \frac{j+\frac{1}{2}}{n} \qquad j=0,\ 1,\ \ldots.,\ n-1 \qquad (3.4)$$

once and for all compute the n different quantiles, q_j, corresponding to the values of p_j in (3.4) by use of

$$q_j = F^{-1}(p_j) \qquad j=0,\ 1,\ \ldots.,\ n-1 \qquad (3.5)$$

This computation can easily be done either by means of an analytical expression for the inverse function, F^{-1}, in (3.5), or by means of numerical methods. The sequence (1.1) can now be generated by the algorithms

$$\begin{aligned} &j:=\text{entier}(n\cdot v_i);\\ &x_i:=q_j \end{aligned} \qquad (3.6)$$

where v_i is generated by (2.9).

To find the correlation, r_x, for a given value of h, we have to compute

$$Ex_i x_{i-1} = \sum_{r=o}^{n-1} \sum_{s=o}^{n-1} q_r \cdot q_s \, P\{x_i = q_r,\ x_{i-1} = q_s\} \qquad (3.7)$$

and then use (3.3), where

$$\begin{aligned} Ex_i &= \frac{1}{n} \sum_{r=o}^{n-1} q_r \\ \text{Var } x_i &= \frac{1}{n} \sum_{r=o}^{n-1} (q_r - Ex_i)^2 \end{aligned} \qquad (3.8)$$

Of course, $P\{x_i = q_r,\ x_{i-1} = q_s\}$ depends on h, and may be a little difficult to find.

To facilitate the computations we can, instead of using the sequence (1.6) in the algorithms (3.6), use a sequence of correlated uniform distributed random variables on the integers 0, 1, 2, ..., n-1

$$w_1,\ w_2,\ \ldots.,\ w_i,\ \ldots \tag{3.9}$$

generated by the following algorithms

$$\begin{aligned}
&w_i := \text{randint}(w_{i-1}-c, w_{i-1}+c);\\
&\underline{\text{if}}\ w_i<0\ \underline{\text{then}}\ w_i:=-1-w_i\ \underline{\text{else}}\\
&\underline{\text{if}}\ w_i\geq n\ \underline{\text{then}}\ w_i:=2n-1-w_i
\end{aligned} \tag{3.10}$$

where randint (a,b) generates a uniform distributed random variable on the integers a, a+1, ..., b-1, b and where c is an integer, and then use

$$\begin{aligned}
&j:=w_i;\\
&x_i:=q_j
\end{aligned} \tag{3.11}$$

For the probability function of w_i and w_{i-1} we have

$$z(w_i,w_{i-1}) = z(w_i|w_{i-1})z(w_{i-1}) \tag{3.12}$$

where (see Fig. 3.1)

$$z(w_i|w_{i-1}) = \left\{\begin{array}{ll}
\frac{1}{2c+1} & \text{if}\left\{\begin{array}{l} c\leq w_i+w_{i-1}\leq n-1-c\\ -c\leq w_i-w_{i-1}\leq c\end{array}\right.\\
\frac{2}{2c+1} & \text{if}\quad w_i+w_{i-1}<c\\
\frac{2}{2c+1} & \text{if}\quad n-1-c<w_i+w_{i-1}\\
0 & \text{else}
\end{array}\right\} \begin{array}{l}0\leq w_i\leq n-1\\ 0\leq w_{i-1}\leq n-1\end{array} \tag{3.13}$$

If the marginal probability function for w_{i-1} is uniform and given by

$$z(w_{i-1}) = \frac{1}{n}\quad \text{for } w_{i-1}=0,\ 1,\ 2,\ \ldots,\ n-1 \tag{3.14}$$

we can deduce that the marginal probability function for w_i is

$$\sum_{k=\max(0,w_i-c)}^{\min(n-1,w_i+c)} z(w_i|w_{i-1})\cdot z(w_{i-1}) = \frac{1}{n} \tag{3.15}$$

$$\text{for } w_i=0,\ 1,\ 2,\ \ldots,\ n-1$$

That is to say the marginal probability function for w_i is uniform, too.

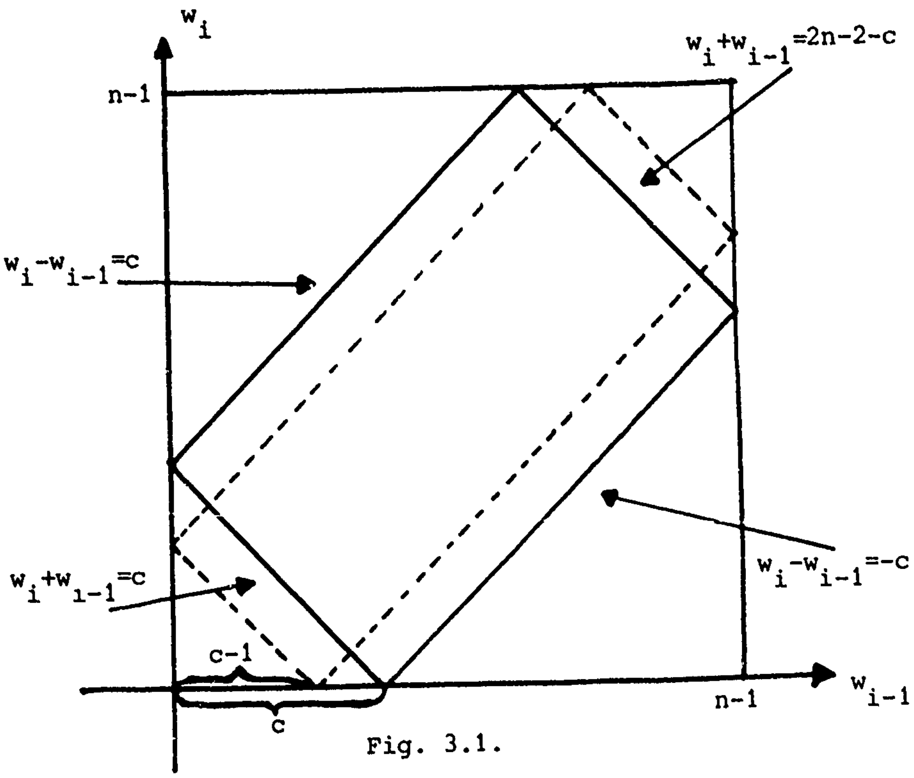


Fig. 3.1.

To find the correlation, r_w, between w_i and w_{i-1} we introduce

$$Ew_iw_{i-1} = \frac{S(c)}{n(2c+1)} \tag{3.16}$$

where (see Fig. 3.1)

$$S(c) = S(c-1) + 2\sum_{k=o}^{n-c-1} k(k+c) + \sum_{k=o}^{c-1} k(c-1-k) + \sum_{k=o}^{c-1} (n-c+k)(n-1-k) \tag{3.17}$$

and where

$$S(0) = \sum_{k=o}^{n-1} k^2 = \frac{(n-1)n(2n-1)}{6} \tag{3.18}$$

From (3.17) we have

$$S(c) - S(c-1) = \frac{1}{3}\left[2c^3 - 3nc^2 + c + n(n-1)(2n-1)\right] \tag{3.19}$$

and hence

$$S(c) = S(0) + \frac{1}{3}\left[\frac{c^2(c+1)^2}{2} - \frac{c(c+1)(2c+1)n}{2} + \frac{c(c+1)}{2} + cn(n-1)(2n-1)\right] \tag{3.20}$$

where $S(0)$ is given by (3.18).

By using

$$Ew_i = \frac{n-1}{2}$$
$$\operatorname{Var} w_i = \frac{n^2-1}{12} \tag{3.21}$$

we have

$$r_w = \frac{12S(c) - 3n(2c+1)(n-1)}{n(2c+1)(n^2-1)} \tag{3.22}$$

If we put

$$c = h \cdot n \tag{3.23}$$

we have from (3.22)

$$\lim_{n\to\infty} r_w = \frac{4\left[\frac{h^4}{2} - h^3 + 2h\right] - 6h}{2h} \tag{3.24}$$

and hence

$$\lim_{n\to\infty} r_w = h^3 - 2h^2 + 1 = r_v \tag{3.25}$$

To find the correlation, r_x for a given value of c, we must compute (3.7), but in this case we have

$$P\{x_i = q_r,\ x_{i-1} = q_s\} = z(q_r, q_s) \tag{3.26}$$

That is to say, we can find $Ex_i x_{i-1}$ almost in the same way as we found $Ew_i w_{i-1}$ in (3.16). To do this we introduce

$$Ex_i x_{i-1} = \frac{R(c)}{n(2c+1)} \tag{3.27}$$

where

$$\begin{aligned} R(c) = R(c-1) + 2 \sum_{k=o}^{n-c-1} q_k q_{k+c} &+ \sum_{k=o}^{c-1} q_k q_{c-1-k} \\ &+ \sum_{k=o}^{c-1} q_{n-c-k} q_{n-1-k} \end{aligned} \tag{3.28}$$

and where

$$R(0) = \sum_{k=o}^{n-1} q_k^2 \tag{3.29}$$

Is is now easy to make an algorithm computing the correlation $r_x(c)$ by means of (3.3), (3.8), (3.27), (3.28) and (3.29), and then find a value of c such that

$$r_x(c) \le r_x \le r_x(c-1) \tag{3.30}$$

where r_x is the value we want to use in the generation of the sequence in (1.1). We can then choose either c or c-1 as the value in the algorithms (3.10).

If we now set h in (2.9) to a value such that h·n is integer, i.e.

$$d = h \cdot n \qquad (d \text{ integer}) \tag{3.31}$$

we can find $P\{x_i = q_r,\ x_{i-1} = q_s\}$ in (3.7) as

$$P\{x_i = q_r,\ x_{i-1} = q_s\} = P\{x_i = q_r | x_{i-1} = q_s\} P\{x_{i-1} = q_s\} \tag{3.32}$$

where

$$P\{x_{i-1} = q_s\} = \frac{1}{n} \tag{3.33}$$

and

$$
P\{x_i = q_r \mid x_{i-1} = q_s\} = \begin{cases} \frac{1}{2}\cdot\frac{1}{2d} & \text{if} \quad -d = r - s \\ \frac{1}{2}\cdot\frac{1}{2d} & \text{if} \quad r - s = d \\ \frac{3}{2}\cdot\frac{1}{2d} & \text{if} \quad d - 1 = r + s \\ \frac{3}{2}\cdot\frac{1}{2d} & \text{if} \quad r + s = n - d \\ \frac{1}{2d} & \text{if} \begin{cases} d \le r + s \le n - 1 - d \\ -d < r - s < d \end{cases} \\ \frac{1}{d} & \text{if} \quad r + s < d - 1 \\ \frac{1}{d} & \text{if} \quad n - d < r + s \\ 0 & \text{else} \end{cases} \left.\vphantom{\begin{matrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\end{matrix}}\right\} \begin{matrix} 0 \le r \le n-1 \\ 0 \le s \le n-1 \end{matrix} \tag{3.34}
$$

That is to say if we introduce

$$
Ex_i x_{i-1} = \frac{Q(d)}{n\cdot 2d} \tag{3.35}
$$

we have

$$
\begin{aligned} Q(d) = Q(d-1) &+ \sum_{k=o}^{n-d-1} q_k q_{k+d} + \sum_{k=o}^{n-d-2} q_k q_{k+d-1} \\ &+ \frac{1}{2}\sum_{k=o}^{d-1} q_k q_{d-1-k} + \frac{1}{2}\sum_{k=o}^{d-2} q_k q_{d-2-k} \\ &+ \frac{1}{2}\sum_{k=o}^{d-1} q_{n-d+k} q_{n-1+k} + \frac{1}{2}\sum_{k=o}^{d-2} q_{n-d-1+k} q_{n-1-k} \end{aligned} \tag{3.36}
$$

where

$$
Q(0) = 0 \tag{3.37}
$$

We are now able to make an algorithm computing the correlation $r_x(d)$ and by means of this algorithm to find, similar to (3.30), a value of d such that

$$r_x(d) \leq r_x \leq r_x(d-1) \tag{3.38}$$

4. ERLANG DISTRIBUTION

For the Erlang distribution we have

$$F(x) = 1 - e^{-sax} \sum_{k=o}^{s-1} \frac{(sax)^k}{k!} \tag{4.1}$$

where

$$Ex = \frac{1}{a}$$

$$\text{Var } x = \frac{1}{s} \frac{1}{a^2} \tag{4.2}$$

Having computed q_j in (3.5), we can find r_x as a function of c by the algorithms mentioned in section 3.1 in connexion with (3.30).

If we introduce

$$e = \frac{c}{n} \tag{4.3}$$

it turns out that $e + r_x$ is very close to 1. The values of $10000 \cdot (e + r_x - 1)$ for different values of s and r_x are shown in tables 4.1, 4.2 and 4.3 for $n = 100$, $n = 200$ and $n = 500$ respectively. By means of these tables we can find by interpolation the value of $e + r_x - 1$ for given r_x and s. Finally, from this value we can find c by (4.3). In Fig. 4.1, 4.2 and 4.3 some of the results are shown. As it can be seen, the values are not very sensitive with respect to changes in n.

Finally, we can find r_x as a function of h by the algorithms mentioned in section 3.1 in connexion with (3.38). The values of $10000 \cdot (h + r_x - 1)$ for different values of n, s and r_x are shown in table 4.4. In Fig. 4.4 some of the results for $h + r_x - 1$ are shown. For comparison the corresponding values for $e + r_x - 1$ from table 4.1 are shown, too.

r_x \ s	1	2	3	4	5	7	10	15	20	50
.05	-257	-136	-100	-82	-72	-61	-52	-46	-43	-37
.10	-400	-184	-118	-86	-67	-47	-31	-20	-14	-4
.15	-499	-206	-115	-71	-45	-16	5	21	29	44
.20	-566	-210	-98	-43	-11	25	51	72	82	100
.25	-609	-201	-70	-6	32	74	105	130	142	163
.30	-631	-181	-34	38	81	129	165	192	206	230
.35	-636	-152	8	87	134	187	227	257	272	299
.40	-627	-118	54	139	190	247	290	323	340	369
.45	-605	-79	101	192	246	307	353	389	406	438
.50	-573	-38	150	245	301	366	415	452	471	505
.55	-531	5	198	296	354	422	472	512	531	567
.60	-480	49	243	343	403	473	525	565	586	622
.65	-423	90	285	385	446	516	570	611	632	670
.70	-359	129	320	419	480	551	604	646	668	706
.75	-289	163	346	443	502	572	625	667	688	727
.80	-215	190	360	451	509	576	628	669	689	727
.85	-139	205	356	440	492	555	604	642	662	698
.90	-61	202	326	397	442	497	539	574	592	624
.95	10	169	251	300	332	372	405	431	444	470

$(e + r_x - 1) \cdot 10000$

$n = 100$

Table 4.1.

r_x \ s	1	2	3	4	5	7	10	15	20	50
.05	-241	-116	-78	-60	-49	-38	-29	-22	-19	-13
.10	-391	-168	-99	-66	-47	-26	-10	2	8	19
.15	-497	-194	-100	-54	-28	2	24	41	50	65
.20	-571	-202	-86	-29	4	42	69	90	101	120
.25	-619	-197	-61	5	4[illegible]	89	121	146	159	181
.30	-646	-180	-28	47	91	141	178	207	221	247
.35	-656	-156	11	93	142	197	238	270	286	314
.40	-652	-125	54	142	195	255	300	334	352	383
.45	-634	-90	98	193	249	313	361	398	417	450
.50	-606	-51	144	243	302	370	420	460	479	514
.55	-567	-12	189	291	353	423	476	517	558	575
.60	-520	28	231	336	399	472	526	569	590	629
.65	-466	67	270	375	439	513	569	613	635	674
.70	-404	103	302	406	470	545	601	646	668	709
.75	-336	134	325	427	490	564	620	664	687	727
.80	-263	158	336	433	494	565	620	663	685	725
.85	-185	171	330	419	475	541	593	635	656	694
.90	-105	168	299	374	422	480	526	564	583	618
.95	-27	138	224	277	311	355	390	418	433	461

$(e + r_x - 1) \cdot 10000$

$n = 200$

Table 4.2.

r_x \ s	1	2	3	4	5	7	10	15	20	50
.05	-232	-104	-65	-47	-36	-24	-15	-8	-5	1
.10	-388	-159	-89	-55	-35	-13	3	15	21	32
.15	-499	-188	-92	-45	-18	13	36	53	62	77
.20	-577	-199	-80	-22	12	51	79	101	112	131
.25	-629	-196	-57	11	51	96	130	156	169	192
.30	-660	-183	-26	51	96	148	186	215	230	256
.35	-674	-161	10	95	145	202	245	277	293	323
.40	-672	-132	51	142	197	259	305	340	358	390
.45	-658	-100	94	191	249	315	364	403	422	456
.50	-633	-64	138	239	300	370	422	463	483	520
.55	-597	-26	180	286	349	422	477	519	540	579
.60	-552	11	221	328	394	469	525	570	592	632
.65	-500	48	257	366	432	509	567	612	635	676
.70	-440	81	287	395	462	539	598	644	667	709
.75	-373	110	308	414	479	556	615	661	684	726
.80	-301	132	317	418	481	555	613	658	681	723
.85	-223	144	309	402	460	530	584	627	650	690
.90	-142	140	276	355	405	467	515	555	575	612
.95	-58	111	201	256	293	339	376	407	423	453

$$(e + r_x - 1) \cdot 10000$$

$$n = 500$$

Table 4.3.

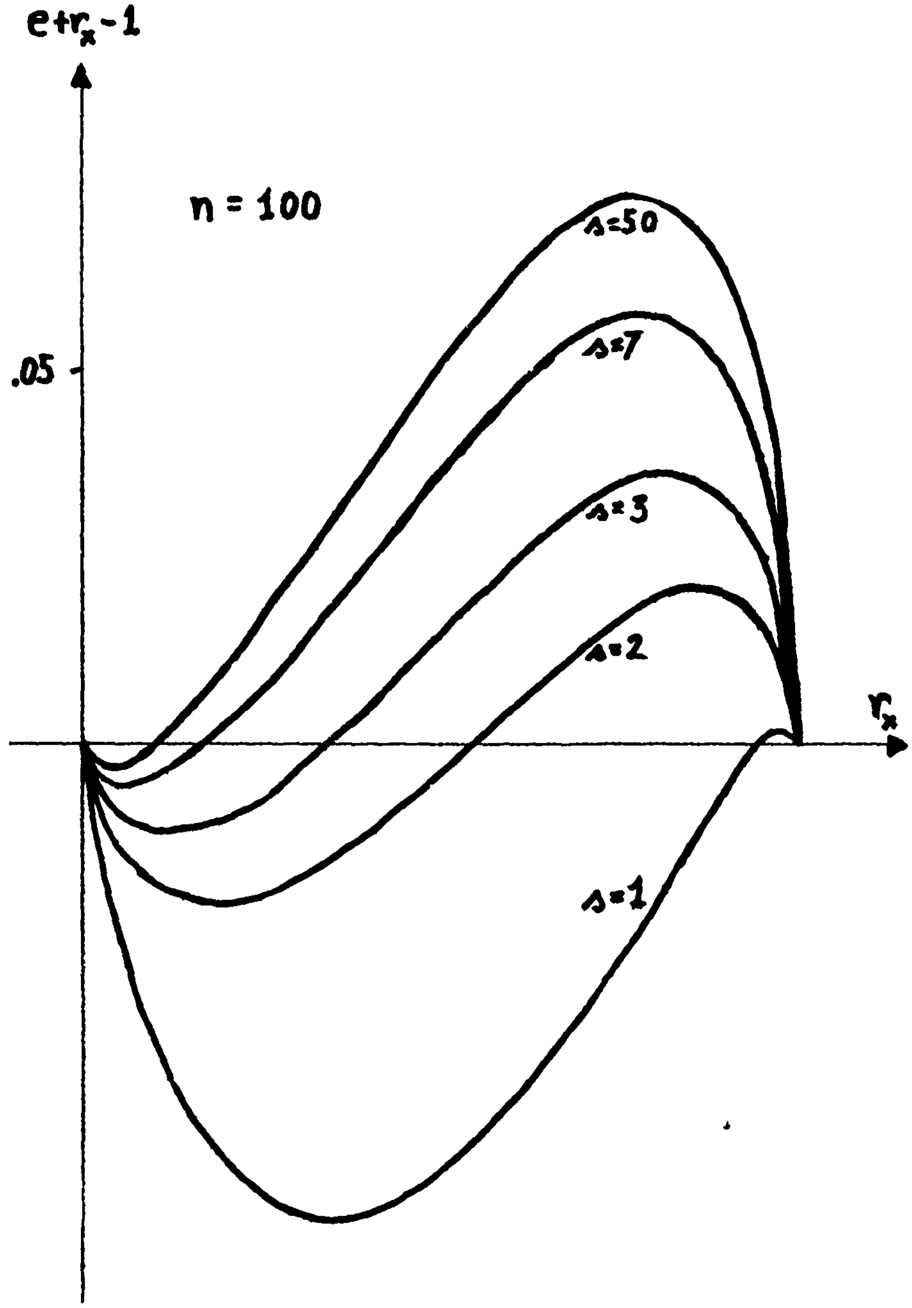


Fig. 4.1.

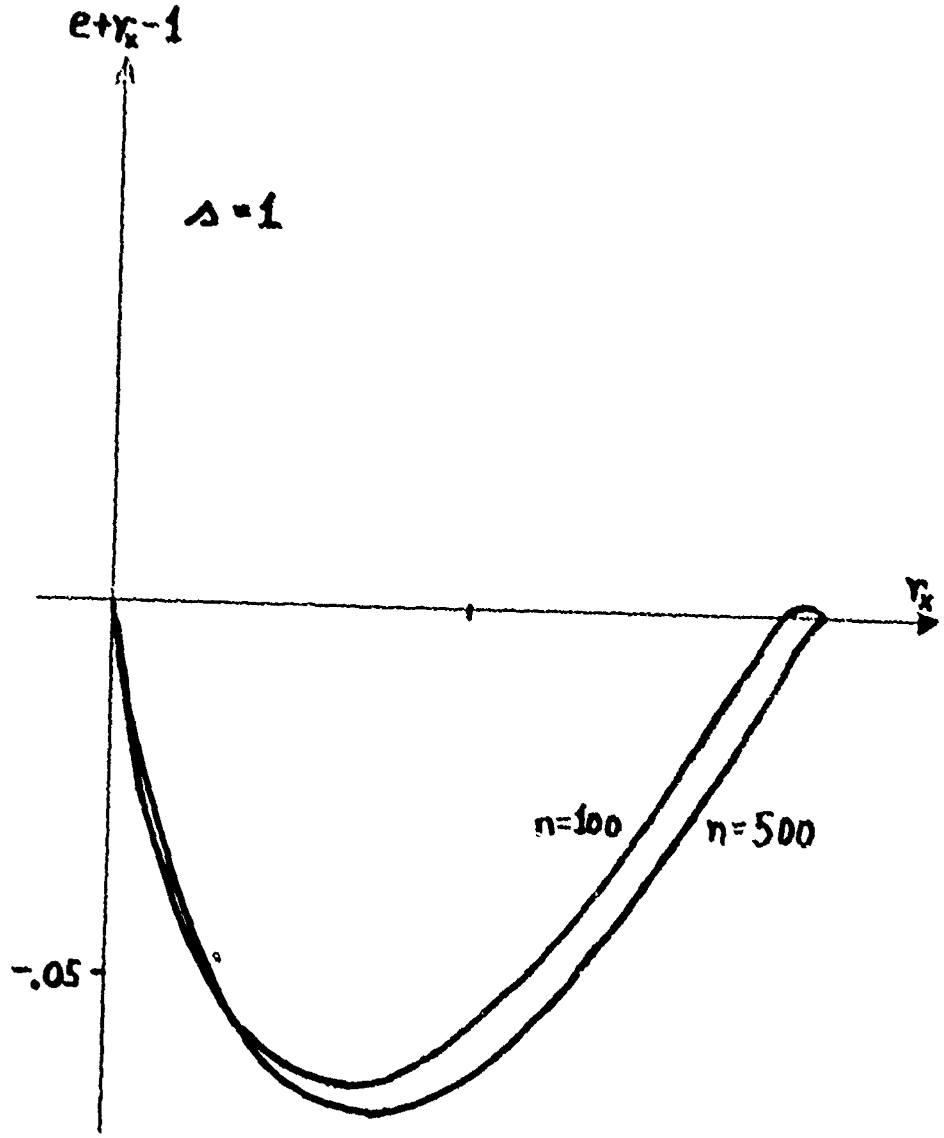


Fig. 4.2.

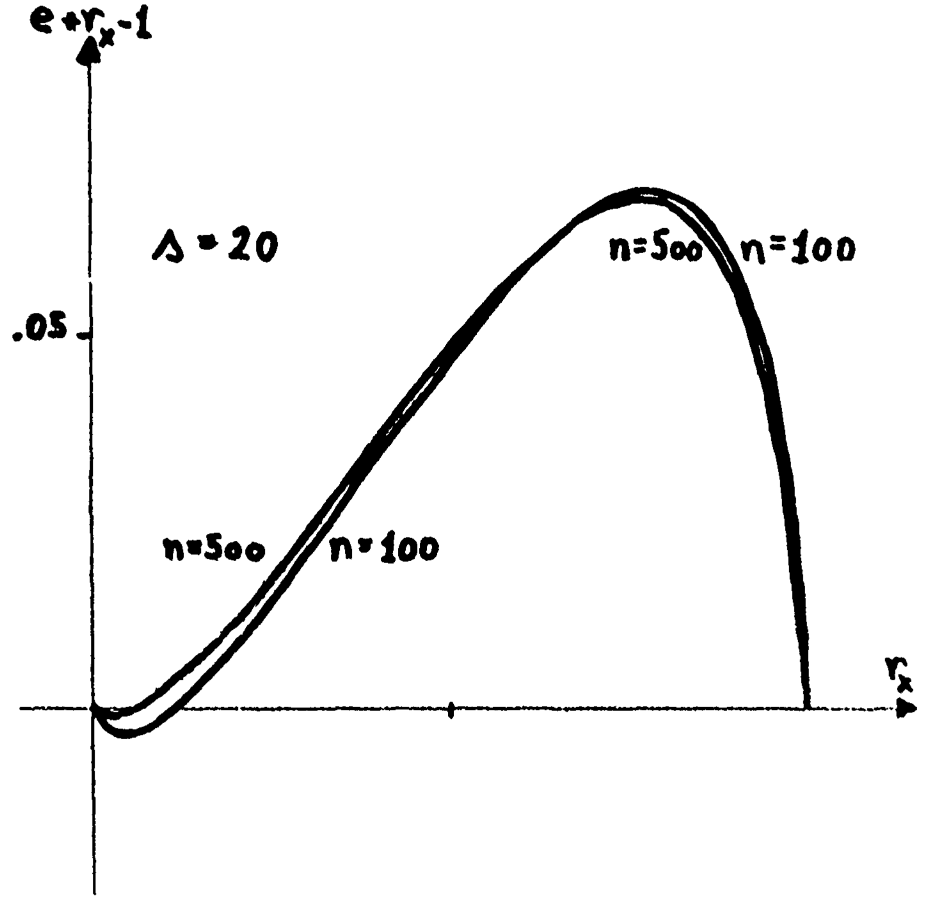


Fig. 4.3.

r_x \ n	100		200		500	
	s = 1	s = 5	s = 1	s = 5	s = 1	s = 5
.05	-207	-22	-216	-24	-222	-26
.10	-350	-18	-366	-22	-378	-26
.15	-449	4	-472	-3	-489	-8
.20	-517	39	-546	29	-567	22
.25	-559	81	-594	69	-619	61
.30	-581	130	-621	116	-650	106
.35	-586	184	-631	167	-664	155
.40	-577	239	-627	220	-662	207
.45	-556	295	-609	274	-648	259
.50	-523	351	-581	327	-623	310
.55	-481	404	-543	377	-587	359
.60	-431	453	-496	424	-542	404
.65	-373	495	-441	464	-490	442
.70	-310	529	-379	495	-430	472
.75	-240	551	-311	515	-363	489
.80	-167	557	-238	519	-291	491
.85	-90	541	-161	499	-213	470
.90	-14	490	-81	447	-132	415
.95	55	379	-3	335	-48	303

$(h + r_x - 1) \cdot 10000$

Table 4.4.

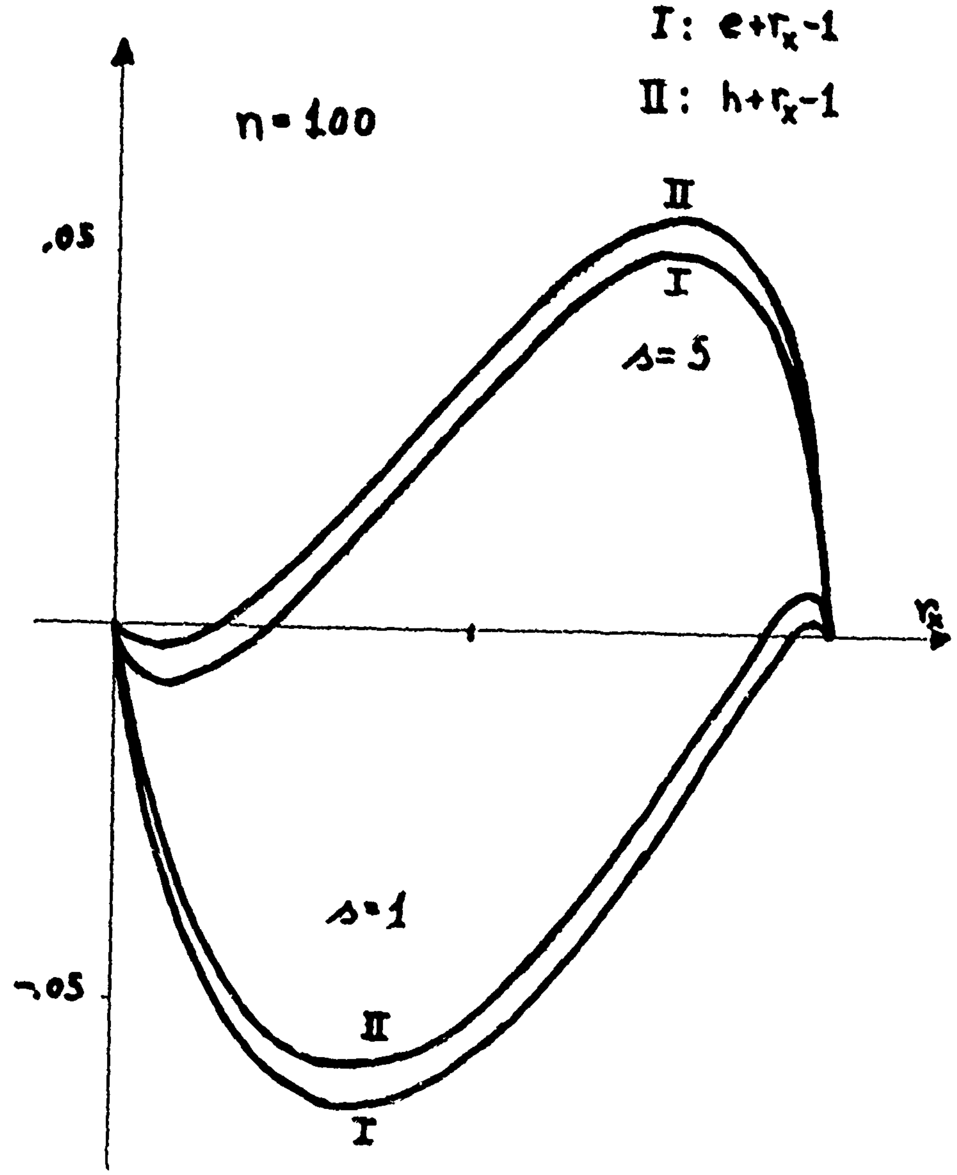


Fig. 4.4.

5. PROCEDURE RANDOM

To compute the quantiles, q_j, in (3.5), the following algol procedure can be used. q_o is lower limit for the quantiles, and delta is a parameter that determines the accuracy of q_j, such that q_j will get a value satisfying

$$\left|\frac{j}{n} + \frac{1}{2n} - F(q_j)\right| < \text{delta} \quad \text{for } j = 0, 1, \ldots, n-1 \tag{5.1}$$

```
procedure random (q,n,qo,delta,F); value n,qo,delta;
array q; integer n; real qo,delta; real procedure F;
begin integer j; real q1,q2,q3,help,p,dt;
      dt:=1/n;
      for j:=0 step 1 until n-1 do
      begin q1:=q3:=qo; p:=(j+.5)/n;
            for q1:=q1+dt while p - F(q1) > 0 do;
            q2:=q1 - dt;
            for qo:=(q1+q2)/2 while true do
            begin help:=p - F(qo);
                  if help > delta then q2:=qo else
                  if help < -delta then q1:=qo else go to e
            end;
e:          q[j]:=qo; dt:=qo - q3
      end
end procedure random;
```

A TECHNIQUE FOR SOLVING LINEAR PROGRAMMING PROBLEMS WITH MULTIPLE OBJECTIVES

MARIO T. TABUCANON and PAKORN ADULBHAN

Division of Industrial Engineering and Management
Asian Institute of Technology
Bangkok, Thailand

ABSTRACT. In recent years, owing to the increasing complexity of contemporary problems, development of multiple-objective optimization models has received significant emphasis. Foremost of all is linear programming which has been the most frequently used.

In this paper, a new method for solving linear programming problems with multiple objectives is introduced. The method is known as the "compromise-constraint technique". It operates in the following manner: To find the compromise solution for any two objectives, a compromise constraint which forces the objectives to be an equal weighted difference from the individual optimal solutions is added to the problem. A single-objective problem with any one of the two original objective functions or their weighted sum, subject to the compromise constraint plus the original ones, is solved. When there are more than two objectives in the problem, the compromise constraints for all possible combinations (by two) of objectives are incorporated, each of which has additional two deviational variables representing positive and negative deviations from the supposed-to-be zero value of the left-hand-side of the standard form of the compromise constraint. The equivalent objective function becomes maximization of a function consisting of two parts -- summation of all weighted objective functions minus (meaning minimization effort) the summation of all deviational variables. A numerical example showing the operation of the compromise-constraint technique is also presented.

1. INTRODUCTION

The need for multiobjective linear programming continues to rise in many decision problems confronting real-world systems. Many decision makers appear to have been evading from this kind of situation due, seemingly, to the handicap faced by the existing state of the art in producing satisfactory results. The evasion is usually done by somehow ignoring secondary objectives (those with lesser degrees of importance) and come up with a single-objective programming model. Apparently, this leads to an erroneous decision.

This paper introduces a solution procedure for solving multiobjective linear programming problems herein known as the "compromise-constraint technique".

2. MATHEMATICAL FORMULATION OF THE PROBLEM

A multiobjective linear programming problem can be generally formulated in the following form:

$$\text{Maximize:} \quad Z_\ell = \sum_{j=1}^{n} c_{\ell j}\, x_j ; \qquad \ell=1, 2,\ldots, k \qquad (1)$$

$$\text{Subject to:} \quad \sum_{j=1}^{n} a_{ij}\, x_j \leq b_i ; \qquad i=1, 2,\ldots, m \qquad (2)$$

$$x_j \geq 0 ; \qquad j=1, 2,\ldots, n \qquad (3)$$

where k (always greater than 1) is the number of objectives; n is the number of decision variables; and m is the number of constraints (excluding non-negativity constraints).

The association of a numerical weight with each objective is imperative in the compromise-constraint technique. Let w_ℓ be the weight associated with objective Z_ℓ. For convenience, weights are made to have the following mathematical traits:

$$0 < w_\ell < 1 ; \qquad \ell=1, 2,\ldots, k \qquad (4)$$

and

$$\sum_{\ell=1}^{k} w_\ell = 1 \qquad (5)$$

The expression $w_\ell > w_{\ell+1}$ denotes that objective Z_ℓ is more important than objective $Z_{\ell+1}$, while the expression $w_\ell = w_{\ell+1}$ indicates their corresponding objectives' equal importance. Determination of these weights are presumed to be made *a priori* and, therefore, beyond the scope of this paper.

3. FRAMEWORK AND METHOD OF SOLUTION

The following principles are the mathematical basis of the compromise-constraint technique.

3.1. Theorem

If $f_\ell(x)$ and $f_h(x)$ are two objective functions whose feasible maxima are Z^*_ℓ and Z^*_h, respectively, and if the functions move toward the feasible region (that is, both functions decrease in numerical value simultaneously), then the equation of the locus of the point or region of intersection of the functions has the following form:

$$\frac{1}{\alpha_\ell}\,[f_\ell(x) - z^*_\ell] - \frac{1}{\alpha_h}\,[f_h(x) - z^*_h] = 0 \qquad (6)$$

where α_ℓ and α_h are the rates at which the respective functions are moving [2].

3.2. Principle

An objective function with bigger weight descends (or relaxes) from its maximum at a slower rate that the objective with smaller weight. In other words, the rates of descent (or rates of relaxation) are inversely proportional to the corresponding weights --

$$\alpha_\ell = K_\ell \frac{1}{w_\ell} \quad \text{and} \quad \alpha_h = K_h \frac{1}{w_h} \qquad (7)$$

where K_ℓ and K_h are constants of proportionality.

Eq. (6) then becomes (after introduction of the values of α_ℓ and α_h and after slight re-arrangement of the terms)---

$$w_\ell\left[\frac{Z_\ell^* - f_\ell(x)}{K_\ell}\right] - w_h\left[\frac{Z_h^* - f_h(x)}{K_h}\right] = 0 \ . \tag{8}$$

Consequently, Eq. (8) becomes the operating equation in looking for the compromise solution; thus, it is referred to, from hereon, as a "compromise constraint". The compromise constraint has to be added to the original constraint set as a means to obtain the compromise solution.

A logical way of getting the compromise solution is by maximizing any one of the individual objective functions -- Z_ℓ and Z_h -- subject to the original set of constraints plus the compromise constraint. Utilization of the weighted sum of the individual objectives, which is defined by the following relationship,

$$Z_{\ell h} = w_\ell \frac{f_\ell(x)}{K_\ell} + w_h \frac{f_h(x)}{K_h} , \tag{9}$$

can also yield the same compromise solution.

For problems with more than two objectives, a zero value of the left-hand-side of a compromise constraints [in a form defined by Eq. (8)] is not necessarily attained. Therefore, addition of deviational variables is imperative. The general formulation of the single-objective equivalent problem of a multiobjective optimization problem using the compromise-constraint technique is as follows:

$$\text{Maximize:} \quad Z = \sum_{\ell=1}^{k} w_\ell \frac{f_\ell(x)}{K_\ell} - \sum_{h\neq\ell} (\sigma_{\ell h}^- + \sigma_{\ell h}^+);$$

$$\ell=1,\ 2, \ldots,\ k; \quad h=1,\ 2, \ldots,\ k \tag{10}$$

Subject to:

$$w_\ell\left[\frac{Z_\ell^* - f_\ell(x)}{K_\ell}\right] - w_h\left[\frac{Z_h^* - f_h(x)}{K_h}\right] + (\sigma_{\ell h}^- - \sigma_{\ell h}^+) = 0;$$

$$\ell=1,\ 2, \ldots,\ k; \quad h=1,\ 2, \ldots,\ k; \quad h\neq\ell \tag{11}$$

$$\sum_{j=1}^{n} a_{ij}\, x_j \leq b_i; \qquad i=1, 2, \ldots, m; \quad (12)$$

$$x_j,\ \sigma^-_{\ell h},\ \sigma^+_{\ell h} \geq 0; \qquad \begin{array}{l} j=1, 2, \ldots, n; \\ \ell=1, 2, \ldots, k; \\ h=1, 2, \ldots, k; \\ h\neq\ell. \end{array} \quad (13)$$

Variable $\sigma^-_{\ell h}$ refers to the negative deviation while $\sigma^+_{\ell h}$ represents the positive deviation from the supposed-to-be zero value of the compromise constraint developed for objectives Z_ℓ and Z_h. On one hand, closest adherence to Eq. (8) for all compromise constraints is aspired for by management. Thus, the expression $(\sigma^-_{\ell h} + \sigma^+_{\ell h})$ has to be minimized; this leads to mutual exclusiveness of $\sigma^-_{\ell h}$ and $\sigma^+_{\ell h}$. On the other hand, management seeks for maximum feasible value of the objective functions. Joining hands together, it appears that the equivalent single objective function has to be the weighted sum of the individual objective functions minus the summation of all deviational variables. The negative sign preceeding the second part of the said function symbolizes the minimization effort for the deviational variables. It is assumed that management gives equal preferences to both parts of the generalized objective function.

The determination of the constants of proportionality in the compromise constraints is rather a complicated affair for a general case. For purposes of this paper, only those for linear objective functions will be considered. Factors $\frac{Z^*_\ell - f_\ell(x)}{K_\ell}$ and $\frac{Z^*_h - f_h(x)}{K_h}$ can be geometrically interpreted as distances of the perpendiculars drawn from any point 'o' (Fig. 1), whose coordinates are defined by (x), to the individual functions -- represented by lines 'oa' and 'ob' in the figure. Point 'o' is a point on the compromise constraint 'cd'. If distances are directly measured from the individual optima, lines 'oa' and 'ob' are also respectively equal in length to lines '$x^{h*}(x^h)$' and '$x^{\ell *}(x^\ell)$'. Any point (x) has a perpendicular distance (D) to a linear equation --

$$\sum_{j=1}^{n} c_{hj}\, x_j = Z^*_h \qquad (14)$$

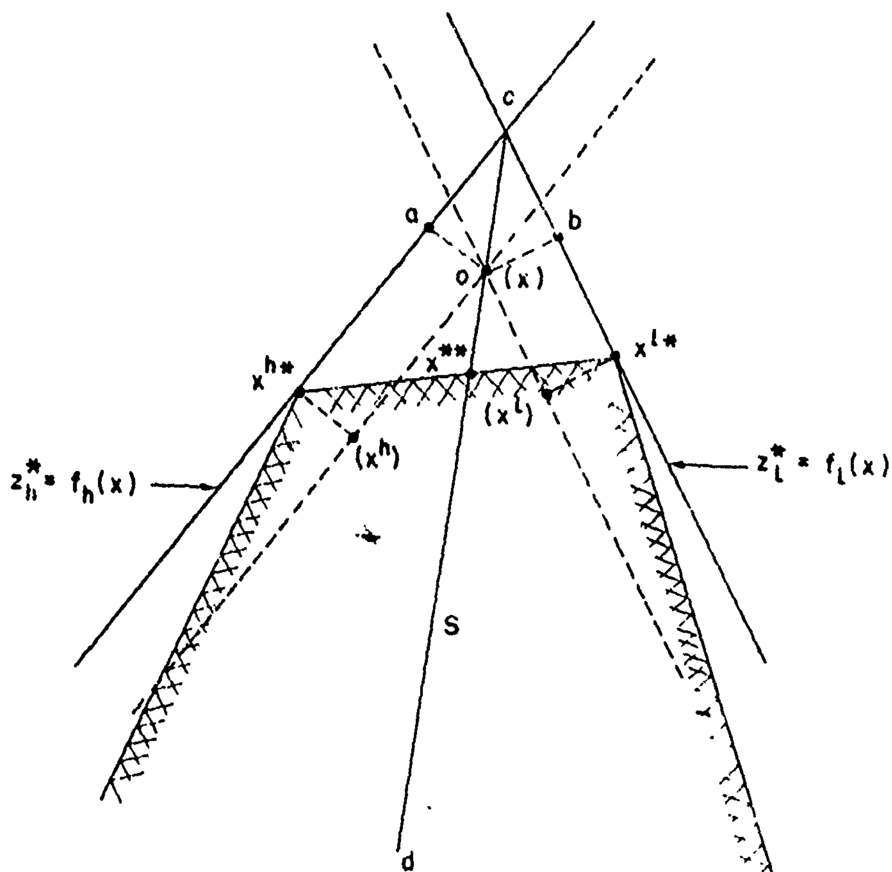


Fig. 1 Graphical Illustration of a Compromise Constraint
Legend : x^{**} Symbolizes a Compromise Solution

-- given by the following expression:

$$D = \frac{\sum_{j=1}^{n} c_{hj}\, x_j - Z_h^*}{\sum_{j=1}^{n} c_{hj}^2} \tag{15}$$

Relating Eq. (15) with Eq. (8) leads to the following relationship:

$$K_h = \sum_{j=1}^{n} c_{hj}^2 \ . \tag{16}$$

Similarly,

$$K_\ell = \sum_{j=1}^{n} c_{\ell j}^2 \ . \tag{17}$$

4. NUMERICAL EXAMPLE

Consider the following three-objective linear programming problem:

$$\text{Maximize:} \quad Z_1 = -x_1 + 3x_2, \; w_1 = 0.1 \tag{18}$$

$$Z_2 = 2x_1 + x_2, \; w_2 = 0.3 \tag{19}$$

$$Z_3 = -2x_1 + x_2, \; w_3 = 0.6 \tag{20}$$

$$\text{Subject to:} \quad -x_1 + x_2 \leq 1 \tag{21}$$

$$x_1 + x_2 \leq 7 \tag{22}$$

$$x_1 \leq 5 \tag{23}$$

$$x_2 \leq 3 \tag{24}$$

$$x_1, x_2 \geq 0 \tag{25}$$

The simplex algorithm yields the following individual optimal solutions:

$$x_1^{1*} = 2, \quad x_2^{1*} = 3, \quad Z_1^* = 7$$

$$x_1^{2*} = 5, \quad x_2^{2*} = 2, \quad Z_2^* = 12$$

$$x_1^{3*} = 0, \quad x_2^{3*} = 1, \quad Z_3^* = 1$$

The three necessary compromise constraints are derived as follows:

(i) For Z_1 and Z_2 --

$$0.1\left[\frac{7 - (-x_1 + 3x_2)}{\sqrt{(-1)^2 + (3)^2}}\right] - 0.3\left[\frac{12 - (2x_1 + x_2)}{\sqrt{(2)^2 + (1)^2}}\right] +$$

$$(\sigma_{12}^- - \sigma_{12}^+) = 0$$

or, $$0.3x_1 + 0.39x_2 + \sigma_{12}^- - \sigma_{12}^+ = 1.388 \; . \tag{26}$$

(ii) For Z_2 and Z_3 --

$$0.3\left[\frac{12 - (2x_1 + x_2)}{\sqrt{(2)^2 + (1)^2}}\right] - 0.6\left[\frac{1 - (-2x_1 + x_2)}{\sqrt{(-2)^2 + (1)^2}}\right] +$$

$$(\sigma_{23}^- - \sigma_{23}^+) = 0 \; .$$

or, $$-6x_1 + x_2 + \sigma_{23}^- - \sigma_{23}^+ = -10 \tag{27}$$

(iii) For Z_1 and Z_3 --

$$0.1\left[\frac{7 - (-x_1 + 3x_2)}{\sqrt{(-1)^2 + (3)^2}}\right] - 0.6\left[\frac{1 - (-2x_1 + x_2)}{\sqrt{(-2)^2 + (1)^2}}\right] +$$

$$(\sigma_{31}^- - \sigma_{31}^+) = 0 \; .$$

or, $$-0.504x_1 + 0.173x_2 + \sigma_{31}^- - \sigma_{31}^+ = 0.0465 \; . \tag{28}$$

The required objective function is derived as follows:

$$Z = 0.1\left[\frac{-x_1 + 3x_2}{\sqrt{(-1)^2 + (3)^2}}\right] + 0.3\left[\frac{2x_1 + x_2}{\sqrt{(2)^2 + (1)^2}}\right] +$$

$$0.6\left[\frac{-2x_1 + x_2}{\sqrt{(-2)^2 + (1)^2}}\right] - \left[\sigma_{12}^- + \sigma_{12}^+ + \sigma_{23}^- + \sigma_{23}^+ + \sigma_{31}^- + \sigma_{31}^+\right]$$

$$\text{or, } Z = -0.3x_1 + 0.497x_2 - \sigma_{12}^- - \sigma_{12}^+ - \sigma_{23}^- - \sigma_{23}^+ - \sigma_{31}^- - \sigma_{31}^+ . \quad (29)$$

In summary, the equivalent single-objective problem of the original three-objective linear programming problem is as follows:

Maximize:

$$Z = -0.3x_1 + 0.497x_2 - \sigma_{12}^- - \sigma_{12}^+ - \sigma_{23}^- - \sigma_{23}^+ - \sigma_{31}^- - \sigma_{31}^+$$

Subject to:

$$\begin{aligned}
0.3x_1 + 0.039x_2 + \sigma_{12}^- - \sigma_{12}^+ &= 1.388 \\
-6x_1 + x_2 + \sigma_{23}^- - \sigma_{23}^+ &= -10 \\
-0.504x_1 + 0.173x_2 + \sigma_{31}^- - \sigma_{31}^+ &= 0.0465 \\
-x_1 + x_2 &\leq 1 \\
x_1 + x_2 &\leq 7 \\
x_1 &\leq 5 \\
x_2 &\leq 3 \\
x_1, x_2, \sigma_{12}^-, \sigma_{12}^+, \sigma_{23}^-, \sigma_{23}^+, \sigma_{31}^-, \sigma_{31}^+ &\geq 0
\end{aligned}$$

The resulting compromise solution has objective values of $Z_1^{**} = 6.83$, $Z_2^{**} = 7.34$ and $Z_3^{**} = -1.34$ at $x_1^{**} = 2.17$ and $x_2^{**} = 3$. Fig. 2 shows the graphical version of the problem and its solution.

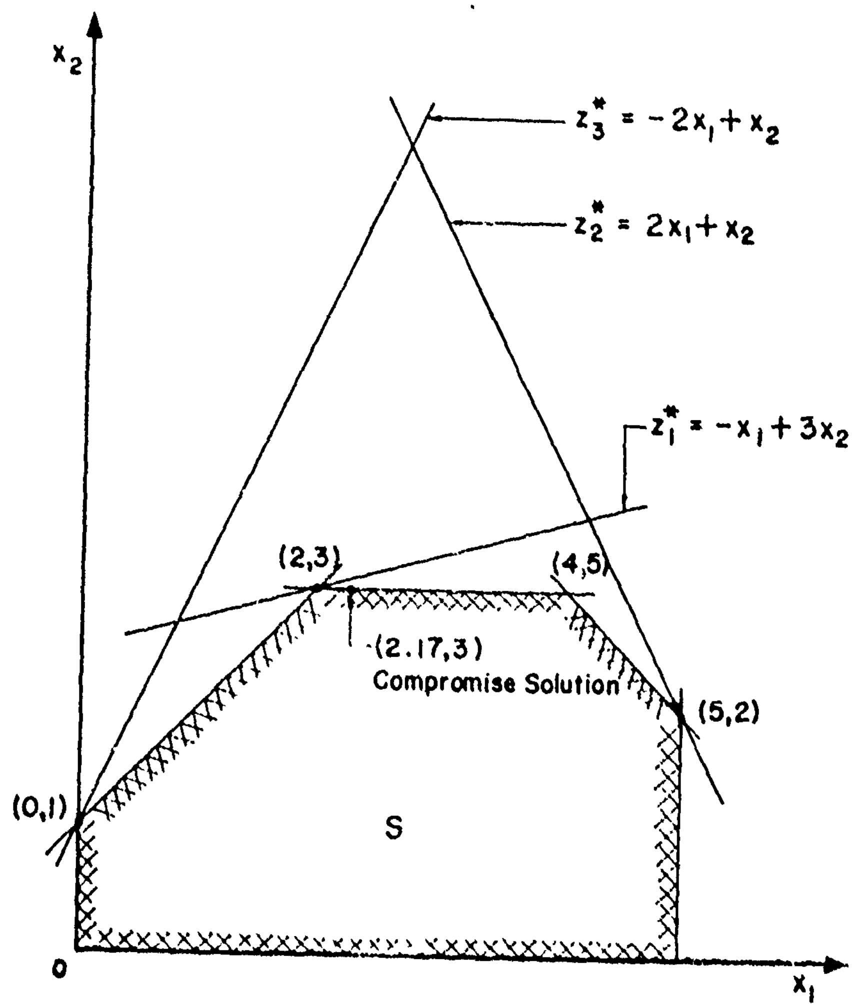


Fig. 2 Graphical Version of Numerical Example

5. CONCLUDING REMARKS

A leading aspect of the compromise-constraint technique is that the compromise solution which is not an extreme point of the feasible region may be found. Given that weights are exogenously determined, a usual way to proceed is to simply derive the weighted sum of the objective functions and solve the resulting single-objective problem. The compromise solution found in this manner for linear programming problems will always be at an extreme point (or more than one). By using the compromise-constraint technique, non-extreme point solutions can be found resulting from the incorporation of compromise constraints.

6. REFERENCES

[1] Tabucanon, M.T., MULTIOBJECTIVE LINEAR PROGRAMMING FOR DECISION MAKING IN INDUSTRIAL SYSTEMS, Doctoral Dissertation, Asian Institute of Technology, Bangkok, 1977.

[2] Adulbhan, P. and Tabucanon, M.T., BICRITERION LINEAR PROGRAMMING, International Journal of Computers and Operations Research, Vol. 4, 1977.

MAXIMUM FLOWS IN A VERTEX WEIGHTED NETWORK

CHAN GIN-HOR

Department of Mathematics
Nanyang University
Singapore

ABSTRACT. The classical network flow problem is to consider maximial flow from the source to the sink in a network. In this paper, we consider a general problem - vertex weighted network flow problem, in which each vertex is assigned with a weight and the flows through each vertex are assumed to be amplified by the weight at the vertex. The notion of vertex weighted network flow problem was first introduced by W.S. Jewell and was later studied by C. Berge, A. Ghouila, K.H.Tan and H.H. Teh. A characterization theorem of a maximum flow in a vertex weighted network has been obtained in this paper.

1. INTRODUCTION

Let $G = \langle V,E \rangle$ be a connected directed graph. A function w from the vertex set V into the set R^+ of positive real numbers is called a weight function. $w(v)$ is called the weight of vertex v. $N = \langle V,E,w \rangle$ is called a vertex weighted graph. Let u_1 and u_2 be the source and sink of the $N = \langle V,E,w \rangle$ respectively. A flow in $N = \langle V,E,w \rangle$ is a function f from the edge set E into the set R of real numbers such that

$$\sum_{e\epsilon\delta^+(v)} f(e) = w(v) \sum_{e\epsilon\delta^-(v)} f(e)$$

for all $v \epsilon V - \{u_1,u_2\}$, where δ^+ and δ^- are incidence functions defined by :

$\delta^+(v) = \{e\epsilon E | v$ is the starting vertex of $e\}$,

$\delta^-(v) = \{e\epsilon E | v$ is the ending vertex of $e\}$,

$\sum_{e\epsilon\delta^+(u_1)} f(e)$ and $\sum_{e\epsilon\delta^-(u_2)} f(e)$ are called the input and output of the flow f. If each edge e is assigned an interval $[0,c(e)]$, called capacity interval, then $N = \langle V,E,w,c \rangle$ is called a vertex weighted network. A flow f is called a feasible flow in $N = \langle V,E,w,c \rangle$ if $0 \le f(e) \le c(e)$, for all $e \epsilon E$. A maximum vertex weighted network flow problem is to find a feasible flow f in which the input of f is maximum.

The general vertex (weighted network) flow problem was first introduced by W.S. Jewell [5] and independently studied by C.Berge [1]. A structure theorem of the feasible flow space of a vertex weighted network was given by K.H. Tan and H.H. Teh [6] in 1972. In this paper, we derive a necessary and sufficient condition for a maximum flow in a vertex weighted network.

2. LOOPS OF A VERTEX WEIGHTED GRAPH.

Let $N = \langle V,E,w \rangle$ be a vertex weighted graph. Suppose $V = \{v_1,v_2,\ldots,v_p\}$, $E = \{e_1,e_2,\ldots,e_q\}$ and $w_i = w(v_i)$, $i = 1,2,\ldots,p$. Define a $p \times q$ matrix $A = (a_{ij})$ as follows :

$$a_{ij} = \begin{cases} 1, & \text{if } e_j \varepsilon \delta^+(v_i), \\ -w_i, & \text{if } e_j \varepsilon \delta^-(v_i), \\ 0\ , & \text{otherwise.} \end{cases}$$

A is called the incidence matrix of $N = <V,E,w>$

Example 2.1. Given a vertex weighted graph N :

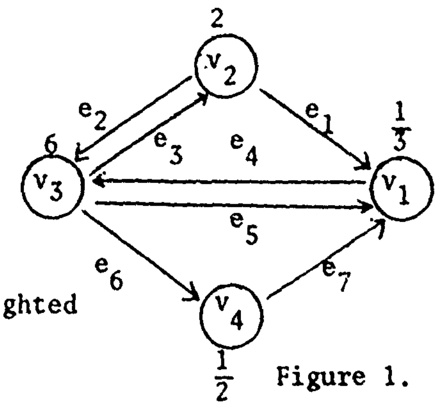


Figure 1.

The incidence matrix of N is

$$A = \begin{pmatrix} -\frac{1}{3} & 0 & 0 & 1 & -\frac{1}{3} & 0 & -\frac{1}{3} \\ 1 & 1 & -2 & 0 & 0 & 0 & 0 \\ 0 & -6 & 1 & -6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 \end{pmatrix}$$

Let $R^q = \overbrace{R \times R \times \ldots \times R}^{q \text{ terms}}$. For each vector $X = (x_1,\ldots,x_q) \varepsilon R^q$, define

$$\begin{aligned} car^+(X) &= \{e_i \varepsilon E \mid x_i > 0\}, \\ car^-(X) &= \{e_i \varepsilon E \mid x_i < 0\}, \\ car\ (X) &= car^+(X)\ car^-(X) \end{aligned}$$

$car^+(X)$, $car^-(X)$ and car (X) are called positive carrier, negative carrier and carrier of X respectively. X is called a loop of $N = <V,E,w>$ if $AX^t = 0$. A loop X is said to be a minimal loop if there exists no loop Y such that car(X) properly contains car(Y).

Remark 2.2. Let $Y = (x_1, x_2,\ldots,x_q)$ be a minimal loop of $N = <V,E,w>$. Then the subgraph H induced by car X has no end vertex.

In fact if v_i is an end vertex of H, we may assume $\delta^+(v_i) = 0$ and $\delta^-(v_i) = \{e_j\}$. Then the i^{th} row of the incidence matrix A of N is $(0,\ldots,0,-w_j,0,\ldots,0)$ and hence

$$(0,\ldots,0,-w_j,0,\ldots,0)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix} = -x_j w_j \neq 0$$

Thus $AX^t \neq 0$, which contradicts to the assumption that X is a loop.

A path P of length n from vertex v_{α_0} to vertex v_{α_n} is defined as the composite concept of a sequence of n+1 distinct vertices $(v_{\alpha_0}, v_{\alpha_1}, \ldots, v_{\alpha_n})$ and a sequence of n edges with signs $(\varepsilon_1 e_{\beta_1}, \varepsilon_2 e_{\beta_2}, \ldots, \varepsilon_n e_{\beta_n})$, where $\varepsilon_i = 1$ or -1, such that, for each i

$$e_{\beta_i} \in \delta^+(v_{\alpha_{i-1}}),\text{ when } \varepsilon_i = 1,$$

$$e_{\beta_i} \in \delta^+(v_{\alpha_i}),\quad\text{ when } \varepsilon_i = 1,$$

$$e_{\beta_i} \in \delta^-(v_{\alpha_{i-1}}),\text{ when } \varepsilon_i = -1,$$

$$e_{\beta_i} \in \delta^+(v_{\alpha_i}),\quad\text{ when } \varepsilon_i = -1.$$

Edge e_{β_i} is said to lie on the path in the positive direction or in the negative direction according to whether $\varepsilon_i = 1$ or -1. P is called a cycle if $v_{\alpha_0} = v_{\alpha_n}$. Note that a path or a cycle can be considered as a subgraph of N. A connected subgraph of N is called a two-cycle if it is composed of two cycles and a path connecting them.

Since a two-cycle H with n vertices has n+1 edges, its incidence matrix A has non-trivial kernel. Thus a loop of H (which in turn is a loop of N) always exists. On the other hand, it is easily seen that for each minimal loop X of N, car(X) induces a subgraph which is either a two-cycle or a cycle.

Now, let X be a vector in R^q. Assume that car(X) contains n edges and induces a subgraph H. Let X_1 be the vector in R^n

obtained by deleting the zero components of X. Let A_1 be the incidence matrix of H. If H is a cycle, then A_1 is an n x n matrix. Thus if X is a loop of N, $A_1X_1^t = 0$ and hence $\det(A_1) = 0$.

Suppose H is written as $(v_{\alpha_0}, v_{\alpha_1}, \ldots, v_{\alpha_n}), (\varepsilon_1 e_{\beta_1}, \varepsilon_2 e_{\beta_2}, \ldots, \varepsilon_n e_{\beta_n})$. Then it is not difficult to see that $\det(A_1) = 0$ if and only if $\underset{\varepsilon_i=1}{\pi} w_{\alpha_i} = \underset{\varepsilon_i=-1}{\pi} w_{\alpha_i}$ or $\overset{n}{\underset{i=1}{\pi}} w_{\alpha_i} = 1$ in the case $\varepsilon_i = 1$ for all i or $\varepsilon_i = -1$ for all i.

We conclude the above-results by the following theorem.

Theorem 2.3. Let $X = (x_1, x_2, \ldots, x_q) \in R^q$ be a minimal loop of $N = \langle V,E,w \rangle$ and suppose that car(X) induces a subgraph H of N. Then we have

(1) H is a cycle or a two-cycle.

(2) If H is a cycle and all the edges of H are in one direction, then the product of the weights of the vertices in H is equal to one.

(3) If H is a cycle and some of the edges in H are in opposite direction, then the product of the weights of the ending vertices of the edges in one direction is equal to the product of the weights of the ending vertices of the edges in the other direction.

Conversely, if a subgraph H of N has property (2) or (3), then there exists a minimal loop X such that car(X) induces H.

Let W be a subspace of R^q and $X,Y \in W$. Denote $X \mathfrak{C} Y$ if

$$\mathrm{car}^+(X) \subseteq \mathrm{car}^+(Y)$$

$$\mathrm{car}^-(X) \subseteq \mathrm{car}^-(Y)$$

The following theorem, due to Farkas [2] is stated without proof.

Theorem 2.4. (Equisignum Decomposition Theorem). Every non-zero vector X in W can be expressed as

$$X = X_1 + X_2 + \ldots + X_s$$

where X_i is minimal in W and $X_i \ll X$.

Corollary 2.5. Every loop X of $N = \langle V,E,w\rangle$ is a sum of minimal loops X_i, $i = 1,2,\ldots,3$, where $X_i \ll X$.

3. FLOWS AND LOOPS. Let $N = \langle V,E,w\rangle$ be a vertex weighted graph with source $u_1 = v_1$ and the sink $u_2 = v_p$. Without loss of generally, we may assume $w(v_1) = w_1 = 1$, $w(v_p) = w_p = 1$. Let $M > \lambda > 0$ be two real numbers. Define a vertex weighted graph S, called (M,λ) graph, as follows :

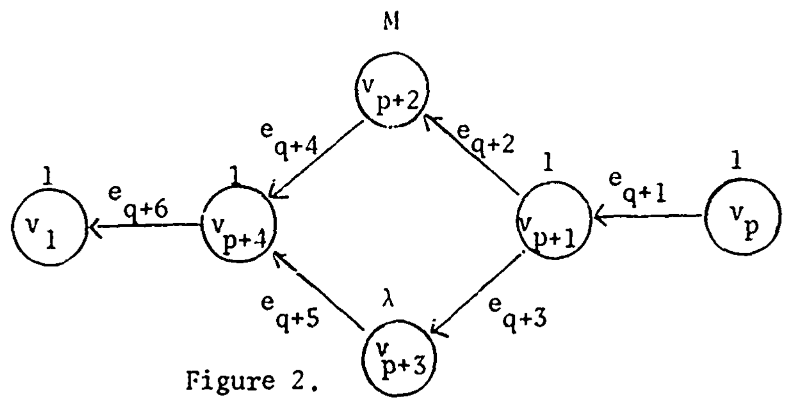


Figure 2.

After joining the source (respectively,the sink) of S to the sink (resp'ly, the source) of N, we get a new vertex weighted graph $N' = \langle V',E',W'\rangle$, called the (M,λ) extended vertex weighted graph of N.

Suppose $f : E \to R$ is a flow in N with $f(e_i) = x_i$. Then x_i is called a flow in edge e_i, $i = 1,2,\ldots,q$, and f can be written as a vector $X = (x_1,x_2,\ldots,x_q)$. Let x and y be the input and output of X respectively. Then the vector

$$X' = (x_1,x_2,\ldots,x_q,y,\frac{x-\lambda y}{M-\lambda},\frac{My-x}{M-\lambda},\frac{M(x-\lambda y)}{M-\lambda},\frac{\lambda(My-x)}{M-\lambda},x)$$

is a loop of N'. On the other hand, if $Y = (y_1,y_2,\ldots,y_{q+6})$ is a loop of N', then $Y' = (y_1,y_2,\ldots,y_q)$ is a flow in N.

Theorem 3.1. A vector $Y = (x_1,x_2,\ldots,x_q)$ is a flow in a vertex weighted graph if and only if

$$X' = \left(x_1, x_2, \ldots, x_q, y, \frac{x-\lambda y}{M-\lambda}, \frac{My-x}{M-\lambda}, \frac{M(x-\lambda y)}{M-\lambda}, \frac{\lambda(My-x)}{M-\lambda}, x\right)$$

is a loop of the (M,λ) extended vertex weighted graph of N where x and y are the input and output of X respectively.

4. CONDUCTIVE PATHS AND CONDUCTIVE CYCLES.

Let $N = \langle V,E,w,c \rangle$ be a vertex weighted network. For a feasible flow $X = (x_1, x_2, \ldots x_q)$ in N, a path $P : (v_{\alpha_0}, v_{\alpha_1}, \ldots v_{\alpha_n}), (\varepsilon_1 e_{\beta_1}, \varepsilon_2 e_{\beta_2}, \ldots, \varepsilon_n e_{\beta_n})$ from v_{α_0} to v_{α_n} is called a conductive path with respect to X if

(1) $d_{\beta_i} = c_{\beta_i} - x_{\beta_i} > 0$, if e_{β_i} lies on P in the positive direction,

(2) $d_{\beta_i} = x_{\beta_i} > 0$, if e_{β_i} lies on P in t e negative direction

d_{β_i} will be called the slack of edge e_{β_i} with respect to X and $D_p = (d_{\beta_1}, d_{\beta_2}, \ldots, d_{\beta_n})$ called the slack of P with respect of X.

For a conductive path P with respect to X given as above, let

$$s_i = \begin{cases} 0 & \text{for } i = 0, \\ \dfrac{\varepsilon_i + \varepsilon_{i+1}}{2}, & \text{for } i = 1,2,\ldots,n-1, \end{cases}$$

$$a_{\alpha_i}(P) = \prod_{j=0}^{i} w_{\alpha_j}^{s_j}, \qquad \text{for } i = 0,1,2,\ldots,n-1,$$

$$b_{\beta_i}(P,D_p) = \frac{d_{\beta_i}}{a_{\alpha_{i-1}}(P)}, \qquad \text{for } i = 1,2,\ldots,n,$$

$$\theta(P,D_p) = \min_i (b_{\beta_i}(P,D_p)),$$

$$\theta_{\beta_i}(P,D_p) = a_{\alpha_{i-1}}(P)\theta(P,D_p), \quad \text{for } i = 1,2,\ldots,n.$$

$\theta(P,D_p)$ will be called the conductive value of the conductive path P. In the case V_{α_0} is the source and v_{α_n} is the sink, we can construct a new feasible flow $Y = (y_1,y_2,\ldots,y_q)$ from $X = (x_1,x_2,\ldots,x_q)$ as follows :

$$y_{\beta_i} = x_{\beta_i} + \varepsilon_i\, \theta_{\beta_i}(P,D_p), \quad i = 1,2,\ldots,n,$$

$$y_j = x_j, \qquad\qquad \text{otherwise.}$$

Clearly, the input of Y is increased by an amount $\theta(P,D_p)$ from that of X. This shows that for a feasible flow X in $N = \langle V,E,w,c\rangle$, if there exists a conductive path from the source to the sink with respect to X, then X is not a maximum flow i.e. the input of X is not maximum.

Example 4.1. Given a vertex weighted network N.

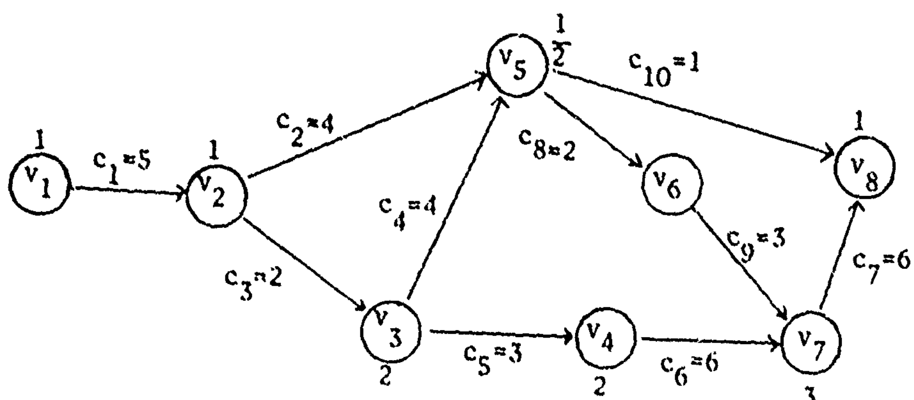


Figure 3.

$X = (1,0,1,2,0,0,0,0,0,1)$ is a feasible flow in N with input 1.

$P = (v_1,v_2,v_5,v_3,v_4,v_7,v_8),(e_1,e_2,-e_4,e_5,e_6,e_7)$ is a conductive path from the source to the sink with respect to X. The slack of P is $D_p = (4,4,2,3,6,6)$. Then we have

i	0	1	2	3	4	5	6
ε_i		1	1	-1	1	1	1
s_i	0	1	0	0	1	1	
w_{α_i}	1	1	½	2	2	3	1
$w_{\alpha_i}^{s_j}$	1	1	1	1	2	3	
$a_{\alpha_i}(P) = \prod_{j=0}^{i} w_{\alpha_j}^{s_i}$	1	1	1	1	2	6	
d_{β_i}		4	4	2	3	6	6
$b_{\beta_i}(P,D_p) = \dfrac{d_{\beta_i}}{a_{\alpha_{i-1}}(P)}$		4	4	2	3	3	1
$\theta_{\beta_i}(P,D_p) = a_{\alpha_{i-1}}(P)\theta(P,D_p)$		1	1	1	1	2	6
$\varepsilon_i\theta_{\beta_i}(P,D_p)$		1	1	-1	1	2	6
x_{β_i}		1	0	2	0	0	0
$y_{\beta_i} = x_{\beta_i} + \varepsilon_i\theta_{\beta_i}(P,D_p)$		2	1	1	1	2	6

where $\theta(P,D_p) = \min(b_{\beta_i}(P,D_p)) = 1$. Hence the vector $Y = (2,1,1,1,1,2,6,0,0,1)$ is a new feasible flow in N with input 2.

On the other hand, if $v_{\alpha_0} = v_{\alpha_n}$, then P becomes a cycle C.

Let $s_n = \dfrac{\varepsilon_n+\varepsilon_1}{2}$ and $a(C) = \prod_{j=1}^{n} w_{\beta_j}^{s_i}$. A conductive cycle C is called a deficient (respectively, balanced, excessive) con-

ductive cycle if $a(C) < 1$ (respectively, $a(C) = 1$, $a(C) > 1$). In the case v_{α_0} is the source and C is a deficient conductive cycle with respect to X, we can obtain a new feasible flow $Y = (y_1, y_2, \ldots, y_q)$ from X as follows :

$$y_{\beta_i} = x_{\beta_i} + \varepsilon_i \theta_{\beta_i}(P, D_p), \qquad i = 1,2,\ldots,n,$$

$$y_j = x_j, \qquad \text{otherwise.}$$

Clearly, the input of Y is increased by an amount $\theta(P,D_p)\cdot(1-a(C))$ from that of X and hence X is not a maximum flow in N.

In the case C is a deficient conductive cycle and there exists a conductive path $P_1 : (v_{k_0}, v_{k_1}, \ldots, v_{k_t})$, $(\varepsilon_1' e_{m_1}, \varepsilon_2' e_{m_2}, \ldots, \varepsilon_t' e_{m_t})$ from the source to a vertex in C say $v_{k_t} = v_{\alpha_0} = v_{\alpha_n}$, with slack $D_{p_1} = (d_{m_1}, \ldots, d_{m_t})$. Let

$$s = \frac{\varepsilon_t' + \varepsilon_1}{2} \text{ and } \Lambda^{\varepsilon_t'}(C, v_{k_t}) = \frac{\theta(P,D_p)(1-a(C))}{w_{k_t}^s}.$$

Then $\Lambda^{\varepsilon_t'}(C, v_{k_t}) > 0$ if and only if C is deficient. Let $D_{p_1}' = (d_{m_1}', \ldots, d_{m_t}')$ be defined by putting

$$d_{m_i}' = d_{m_i}, \qquad i = 1,\ldots,t-1,$$

$$d_{m_t}' = \min(d_{m_t}, \Lambda^{\varepsilon_t'}(C, v_{k_t})).$$

We can then construct a new feasible flow $Y = (y_1, y_2, \ldots, y_q)$ from $X = (x_1, x_2, \ldots, x_q)$ as follows :

$$y_{m_i} = x_{m_i} + \varepsilon_i' \theta(P_1, D_{p_1}') a_{m_{i-1}}(P_1), \qquad i = 1,2,\ldots,t,$$

$$y_{\beta_i} = x_{\beta_i} + \frac{\varepsilon_i w^s_{k_t} \theta(P_1, D'_{p_1}) a_{m_{t-1}}(P_1)}{1 - a(C)} a_{\beta_{i-1}}(P), \ i = 1, 2, .., n,$$

$$y_j = x_j, \qquad \text{otherwise.}$$

The input of Y is increased by an amount $\theta(P_1, D'_{P_1})$ from that of X.

Theorem 4.2. Let $N = \langle V,E,w,c\rangle$ be a vertex weighted network. Let $X = (x_1, x_2, \ldots, x_q)$ be a feasible flow in N. Then X is a maximum flow if and only if N has neither a conductive path from the source to the sink nor a conductive path (or trivial path) from the source to a vertex on a deficient conductive cycle with respect to X.

Proof. We have shown the necessity of the theorem. It remains to prove the sufficiency of the thoerem. Let $M > \lambda > 0$ be such that

$$M > \max\{ \frac{1}{\prod_{i=1}^{n} w_i} \mid n \text{ is arbitrary positive integer} \le p\}$$

$$\lambda \le \min\{ \frac{1}{\prod_{i=1}^{n} w_i} \mid n \text{ is arbitrary positive integer} \le p\},$$

Let N' be the (M,λ) extended vertex weighted graph of N and $[-\infty,\infty]$ the capacity interval of each edge in the (M,λ) graph Then $X' = (x_1, x_2, \ldots, x_{q+6})$ is a loop of N' where x_{q+6} and x_{q+1} are the input and output of X respectively and

$$x_{q+2} = \frac{x_{q+6} - \lambda x_{q+1}}{M - \lambda}, \qquad x_{q+4} = \frac{M(x_{q+6} - \lambda x_{q+1})}{M - \lambda},$$

$$x_{q+3} = \frac{Mx_{q+1} - x_{q+6}}{M - \lambda}, \qquad x_{q+5} = \frac{\lambda(Mx_{q+1} - x_{q+6})}{M - \lambda},$$

Now, suppose the contrary, there exists a loop $Y = (y_1, y_2, \ldots, y_{q+6})$ of N' such that $y_{q+6} > x_{q+6}$ and $0 \le y_i \le c_i$, $i = 1,2,..,q$. Consider the equisignum decomposition of the loop $Z = Y - X'$ into minimal loops, i.e.

$$Z = Y - X' = Z^{(1)} + Z^{(2)} + \ldots + Z^{(n)}$$

where $Z^{(i)} \ll Z$, $i = 1,2,\ldots,n$. Then some $Z^{(i)}$, say $Z^{(1)} = (z_1^{(1)},\ldots,z_{q+6}^{(1)})$ should have a positive $z_{q+6}^{(1)}$.

If $z_{p+1}^{(1)} > 0$, then by the choice of M and λ, $z_k^{(1)} > 0$ for $k = q+1,q+2,\ldots,q+6$, and therefore the edges in car$(Z^{(1)})$ not in the (M,λ) graph form a path from the source to the sink. Since $z_k^{(i)}$, $i = 1,2,\ldots,n$ have the same sign for a fixed k and

$$0 \le y_k \le c_k, \qquad 0 \le x_k \le c_k,$$

for all $k \neq q+1,\ldots,q+6$, we have

$$z_k^{(1)} \le y_k - x_k \le c_k - x_k, \qquad \text{if } z_k^{(1)} > 0,$$

$$0 - x_k \le y_k - x_k \le z_k^{(1)}, \qquad \text{if } z_k^{(1)} < 0.$$

Hence the path is conductive with respect to X, which contradicts the assumption of the theorem.

If $z_{p+1}^{(1)} = 0$, then the edges of car$(Z^{(1)})$ not in the (M,λ) graph form a cycle C with a path P (may be trivial) from the source to the cycle C. By the similar arguement as above. The path P and the cycle C are all conductive with respect to X. It is not difficult to see that C is deficient. Hence we have a deficient conductive cycle and a conductive path from the source to the cycle with respect to X. This contradicts our assumption again. Hence X is a maximum flow in N. This completes the proof.

5. FLOWS WITH MAXIMUM OUTPUT.

In the foregoing sections, we concentrate on the problem of feasible flow with maximum input in a vertex weighted network. However, by performing the simultaneous replacement of terms such as

input $\leftrightarrow$ output
deficient cycle $\leftrightarrow$ excessive cycle
path from the source to a vertex $\leftrightarrow$ path from a vertex to the sink

in Theorem 4.2, we can also obtain a necessary and sufficient condition for a feasible flow of maximum output.

REFERENCES

[1] C. Berge and A. Ghouila, HOUSI, PROGRAMMES, JEUX ET RESEAUX DE TRANSPORT, Dunod, Paris, 1962.

[2] J. Farkas, UBER DIE THEORIE DER EINFACHER UNGLEICHUNGEN, Journal fur die reine und angewandte mathematik 124(1902) 1-27.

[3] L.R. Ford and D.R. Fulkerson, FLOWS IN NETWORKS, Princeton University Press, 1962.

[4] M. Iri, NETWORK FLOW, TRANSPORTATION AND SCHEDULING, Academic Press, 1969.

[5] W.S. Jewell, OPTIMAL FLOWS THROUGH NETWORKS WITH GAINS, Inter. Rep. M.I.T.,8 1958; min. of the 2nd Int. Conf. on Oper. Res. Aix-en-Prouince, 1960.

[6] K.H. Tan and H.H. Teh, ON FLOWS IN FINITE WEIGHTED GRAPHS, Nanta Mathematica Vol. V. No.2(1972), 70-83.

DECENTRALIZED APPROACH TO AGC OF TWO-AREA INTERCONNECTED POWER SYSTEM

S. GLANKWAMDEE

FACULTY OF ENGINEERING

KING MONGKUT'S INSTITUTE OF TECHNOLOGY

Bangkok, Thailand.

ABSTRACT In a large scale system like interconnected power system, multi-objective function can-not be avoided. If one area is considered as a subsystem, each area with individual decision makers is allowed to have its own - objective function. Differential game is relevant to define optimallity of the solution. Nash equilibrium solution will be derived for AGC of two-area interconnected power system.

1. INTRODUCTION

In the operation of interconnected power systems, one of the important control problems is the load and frequency control, more recently termed the 'automatic generation control' (AGC). To maintain the area generation-consumption power balance, an AGC responds to variations of frequency and power exchange over inter-area tie-lines and adjusts the outputs on regulating units. The purpose of this paper is to present a new method for decentralized design of AGC for interconnected power systems. The method is based on the differential games theory and the constrained linear regulators designed for individual power area. A strong motivation for developing this method comes from two important facts. First, in the conventional AGC design the same 'one control per area' principle is used. Second, in the modern multicriteria AGC system with centralized design creates conceptual and computational difficulties. Centralized informatons have to be obtained from power areas which are spread over large geographic territories. In our decentralized method, the AGC regulators are designed using the individual area models and applying the concept of autonomous local control to each area and the associated tie-lines.

The decentralized AGC scheme involves decomposition of the interconnected power system into a number of subsystems so that each subsystem represents an area and all tie-lines originating from that area. Each area consists of its own objective functions and constraints. The problem then becomes a multicriteria problem with multidecision makers. This is where differential games theory [1,2] is relevant to define 'optimality'. Once we define optimality in term of rationality, a decentralized area control of proportional plus integral type [3,4] can be designed as a linear quadratic regulator with constraint output feedback. It retain the area control concept and load distribution property of the conventional regulators while, at the same time, guarantees improved system transients and stability margins for the overall interconnected power system.

2. A DECENTRALIZED POWER SYSTEM DYNAMIC MODEL

A linearized dynamic model of a two-area interconnected power sy tem can be described by the equations

$$\dot{x}(t) = Ax(t) + Bu(t) + Dw(t)$$
$$y(t) = Cx(t), \tag{1}$$

where x,u,z,y, are vectors of states, inputs unmeasurable disturbances and outputs, respectively, w(t),x(t) are zero mean mutually independent stationary Guassian noise with corvarianceΛ(t). The mode(1) is composed of two area models and the associated tie-lie models, which can be presented in a decomposed form as

$$\dot{x}_i(t) = A_i x_i(t)+a_{ti}\Delta p_{ei}(t)+B_i u_i(t)+F_i z_i(t)+D_i w_i(t)$$

$$\Delta p_{ei}(t)=\sum_{\substack{j=1\\ j\neq i}}^{2}\{m_{ij}^T x_i(t)+m_{ji}^T x_j(t)\}$$

$$y_i(t) = C_i x(t) \qquad i=1,2 \tag{2}$$

In (2),x_i,u_i,z_i,y_i are vectors of the i-th area states,inputs, disturbances, and outputs, respectively, and ΔP_{ei} is the variation of the total power exchange of the i-th area, such that

$$\Delta p_{ei} = \sum_{\substack{j=1\\ j\neq i}}^{2}\Delta p_{ij}, \qquad i=1,2 \tag{3}$$

where Δp_{ie} are power exchange deviations between the i-th and the j-th area expressed in per unit. For further explanations of the model see [5].

To design a decentralized AGC regulator, the decoupled model (2) is rearranged. In order to achieve a local control law which is linear, independent of the area load disturbance and the steady-state errors of the frequency and tie-line exchange are zero, it is necessary to augment the model (2) with the new state defined as and intergral of the dynamic 'area control error' ACE y_i^o,

$$v_i(t) = v_{io}+\int_{t_o}^{t} y_i^o(t)dt \qquad i=1,2 \tag{4}$$

where $y_i^o = \Delta p_{ei}+b_{si}\Delta f_i$ b_{si} is the bias factor and Δf_i is the frequency deviation in the i-th area. The two-area interconnected power system model (2) can be rewritten as

$$\dot{x}(t) = Ax(t)+B_1u_1(t)+B_2u_2(t)+Fz(t)+Dw(t)$$
$$\dot{v}(t) = Hx(t) \tag{5}$$

where $x(t_o) = N(0,\Sigma(0))$, $v(t_o) = v(0)$.

We see that there is an 'overlapping' of the subsystem state vectors due to the fact that each subsystem in (5) incorporates the corresponding tie-lie models.

3. REGULATOR DESIGN

The trend in the utility industry is strongly to digital control, using the digital computer for calculating generation change etc. a discrete formula of this problem would thus seem of more practical interest. The decentralized model (5) can be discretized with appropriate sampling interval. A linear discrete time of two-area interconnected power system can be described by the equation

$$\begin{aligned} x(k+1) &= Ax(k)+B_1u_1(k)+B_2u_2(k)+Fz(k)+\Gamma w(k) \\ y_i(k) &= C_i x(k) \qquad i=1,2 \end{aligned} \tag{6}$$

where the vectors are as described in the continuous model.

We require that the AGC regulator should be of a decentralized type with linear feedback, then

$$u_i(k) = K_i(k)y_i(k) \qquad i=1,2 \tag{7}$$

where $K_i(k)$ is determined by minimizing the cost function

$$\mathcal{J}_i = x^T(N)Q_i(N)x(N)+\sum_{k=0}^{N-1}\{x^T(k)Q_i(k)x(k)+u_i^T(k)R_i(k)u_i(k)\} \qquad i=1,2 \tag{8}$$

where $Q_i(k)$ is a symmetric nonnegative definite matrix and $R_i(k)$ is a positive definite matrix.

When the decision of each area can not be made independently, a rational have to be considered. It is natural to choose Nash rational for this problem, since decision maker of each area try to do the best for his area assuming that the other area is doing the same.

Lemma 1. If a linear system described by (6) is controlled using a linear control policy (7) then the expected cost (8) can be expressed as

$$E\{\mathcal{J}_i(k)\} = E\{x^T(k)S_i(k)x(k)\}+\sum_{j=k+1}^{N}\mathrm{tr}\ S_i(j)D\Lambda(j-1)D^T \qquad i=1,2 \tag{9}$$

where

$$\begin{aligned} S_i(k) &= Q_i(k)+C_i^TK_i^TR_i(k)K_iC_i \\ &\quad +(A+B_1K_1C_1+B_2K_2C_2)^TS_i(k+1)(A+B_1K_1C_1+B_2K_2C_2) \\ S_i(N) &= Q_i(N) \qquad i=1,2 \end{aligned} \tag{10}$$

<u>Proof</u> Define $\Sigma(k) = E\{x(k)x^T(k)\}$

then
$$\Sigma(k+1) = (A+B_1K_1C_1+B_2K_2C_2)\Sigma(k)(A+B_1K_1C_1+B_2K_2C_2)^T+D\Lambda(k)D^T \tag{11}$$

Eq. (9) and (10) follow by induction.

The necessary condition for Nash equilibrium of this problem are the derivatives of the expected cost-to-go at stage k with respect to each element of $K_i(k)$ equal to zero, for i=1,2 simultaneously. Thus we have

$$K_i^*(k) = M_i(k)(A+B_jK_j(k)C_j)P_i(k)$$

$$i=1,2;\ j=1,2;\ i\neq j \tag{12}$$

or
$$K_i^*(k) = G_i(k)AT_i(k) \qquad i=1,2 \tag{13}$$

where
$$M_i(k) = -\{R_i(k)+B_i^TS_i(k+1)B_i\}^{-1}\ B_i^TS_i(k+1)$$

$$P_i(k) = \Sigma(k)C_i^T\{C_i\Sigma(k)C_i^T\}^{-1}$$

$$G_i(k) = \{I-M_i(k)B_jM_j(k)B_i\}^{-1}\{M_i(k)+M_i(k)B_jM_j(k)\}$$

$$T_i(k) = \{P_i(k)+P_j(k)C_jP_i(k)\}\{I-C_iP_j(k)C_jP_i(k)\}^{-1}$$

$$i=1,2;\ j=1,2;\ i\neq j$$

<u>Theorem 1</u> The output constrained feedback Nash Equilibrium sequence $\{K_i(k)\}$ i=1,2; k=0,1,...,N-1 for the i-th subsystem are given by equation (13) where it is assumed that the inverse of the required matrices exist.

The sequences $\{K_i(k)\}$ i=1,2; k=0,1,...,N-1 of the i-th area are the solutions to a discrete two-point boundary value problem. As with most algorithms of this type it is expected that the convergence condition depends on the initial guess.

4. CONCLUSIONS

This paper gives a rational attempt to the development of a new decentralized linear regulator approach for AGC of interconnected power systems. The method is based on autonomous area control that used the control feedback only measurements from its own area. There is a subtle interplay between the choice of the performance index for each area and the nature of their interactions. The emphasis is placed

on a systematic analytic design method. At this time we do not know how well the method will perform. Additional effort is needed to explore the numerical design for this problem.

5. REFERENCES

1. Starr, A.W., and Y.C.Ho, NONZERO-SUM DIFFERENTIAL GAMES, JOTA. 3,PP. 184-206, 1969.
2. Starr, A.W., and Y.C. Ho, FURTHER PROPERTIES OF NONZERO-SUM DIFFERENTIAL GAMES, JOTA. 3,PP,207-219, 1969.
3. Ermer, C.M., and V.D. Vandelinde, OUTPUT-FEEDBACK GAINS FOR A LINEAR-DISCRETE STOCHASTIC CONTROL PROBLEM, IEEE Trans. Auto. contr., AC-18, PP. 154-157, 1973.
4. Smith, H.W., and E.J. Davidson, DESIGN OF INDUSTRIAL REGULATORS: INTEGRAL FEEDBACK AND FEEDFORWARD CONTROL, Proc. IEE, Vol.199, No.8,1972.
8. Calovic, M., DYNAMIC STATE SPACE MODELS OF ELECTRIC POWER SYSTEMS, Dept. of Electrical and Mechanical Engineering Report, University of Illinois, Urbana, 111, 61801, USA, 1971.

Deformation Method Using Parametric Approach for Solving Nonlinear Programming Problems

HISASHI MINE and YONG-JOON RYANG

Department of Applied Mathematics and Physics

Faculty of Engineering, Kyoto University

Abstract

This paper proposes a deformation method for solving nonlinear programming problems, illustrating a system of structural optimization problems. If an optimal solution of one of these problems is available, then the other problems can be easily solved parametrically by the deformation method with the initial problem whose solution is known. Furthermore, this deformation method is advantageous in that it is likely to obtain the optimal solution even if the initial point is not necessarily in a small neighborhood of the solution. The proposed method is also useful for solving more than one structural optimization problems of the same type.

1. INTRODUCTION

This paper proposes a deformation method for solving practical nonlinear programming problems. Utilizing the nonlinear parametric programming technique [12], the method solves the problem by imbedding it into a suitable one-parameter family of problems.

The approach discussed in this paper was originally developed with the aim of solving a system of structural optimization problems which frequently appears in various kind of engineering design. It is assumed that we have to solve more than one structural problem of the same type.

If an optimal solution of one of these problems is available, then the optimal solutions of the other problems can be easily obtained by using this known problem and its optimal solution as the initial problem of our parametric deformation method.

A numerical example taken from structural optimization problem is given as an illustration of the proposed method.

The method of nonlinear programming does not generally converge to the optimal solution from an arbitrary starting point if the initial estimate is not sufficiently close to the solution. On the other hand, the deformation method described in this paper is advantageous in that it is likely to obtain the optimal solution even if the initial point is not necessarily in a small neighborhood of the solution.

Section 2 describes nonlinear parametric programming problem imbedded into a one-parameter family of problems. Section 3 proposes two algorithms for finding the optimal solution together with their respective flow charts.

Section 4 discusses a nonlinear programming model for optimization of the design of a beam and gives the computational results of this model.

2. FORMULATION OF THE PROBLEM

The nonlinear programming problem to be solved is assumed to have the form :

$$\begin{aligned} &\text{minimize} && f_1(x) \\ &\text{subject to} && g_1(x) = 0, \quad a \le x \le b, \end{aligned} \tag{1}$$

where x is an n-dimensional vector with components $x_1, x_2, \ldots, x_n$. The functions $f_1 : R^n \to R$ and $g_1 = (g_1, g_2, \ldots, g_m)$: $R^n \to R^m$ are twice continuously differentiable in x.

We suppose that the vector x^0 is known to be an optimal solution of the problem :

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && g_0(x) = 0, \quad a \le x \le b, \end{aligned} \tag{2}$$

where f_0 and g_0 have the same dimension as $f_1(x)$ and $g_1(x)$ in (1). In practice, the solution of (2) may be considerably easier to obtain than that of (1).

The parametric problem we consider here is

$$\begin{aligned} &\text{minimize} && (1-e^{-t})f_1(x) + e^{-t}f_0(x) \\ &\text{subject to} && (1-e^{-t})g_1(x) + e^{-t}g_0(x) = 0. \quad a \le x \le b, \end{aligned} \tag{3}$$

where t is a scalar parameter. If we have inequality constraints, they can be transformed into equality constraints by introducing slack variables. In order to simplify notation, we write the objective function and the constraint functions as $f(x,t)$ and $g(x,t)$, respectively. More specifically, $f(x,t) = (1-e^{-t})f_1(x) + e^{-t}f_0(x)$ and $g(x,t) = (1-e^{-t})g_1(x) + e^{-t}g_0(x)$, i.e., problem (3) is rewritten as : For $t \in T$,

$$\begin{aligned} &\text{minimize} && f(x,t) \\ &\text{subject to} && g(x,t) = 0, \quad a \le x \le b, \end{aligned} \tag{4}$$

where $T = [0,+\infty)$.

Starting with parameter value $t = 0$ and the initial vector x^0, the problem (4) may be changed to its original form (1) by a continuous variation of the value of t from 0 to ∞.

Therefore, an optimal solution of (1) may be obtained by solving (2) parametrically as t increases until $t = \infty$ starting with initial condition $x = x^0$ when $t = 0$.

3. DEFINITIONS OF THE REDUCED PARAMETRIC PROBLEM

In this section, we apply the implicit function theorem to problem (4), and obtain some results which may be viewed as an immediate generalization of those given in [11 For problem (1), we adopt the following nondegeneracy assumption : For any $t \in T$ and for any feasible solution x, i.e., $g(x,t) = 0$ and $a \le x \le b$, there exists a partition of x into y and z, where y is m-dimensional and z is (n-m)-dimensional, such that the $m \times m$ matrix $\nabla_y g(y,z,t)$ is nonsingular and $a \le y \le b$.

The components of the vectors y and z are called basic and nonbasic variables, respectively. With respect to the partition $x = (y,z)$, the matrices $\nabla_x f(x,t)$ and $\nabla_x g(x,t)$ are partitioned as $[\nabla_y f(x,t), \nabla_z f(x,t)]$ and $[\nabla_y g(x,t), \nabla_z g(x,t)]$, respectively.

Given $\bar{t} \in T$, let $\bar{x}$ be a feasible solution of (4). Then by the nondegeneracy assumption, the implicit function theorem assures that we can find a partition $x = (y,z)$ and a twice continuously differentiable function $h: R^{n-m} \times T \to R^m$ implicitly determined by solving the nonlinear equation

$$g(y,z,t) = 0$$

for all (z,t) in some neighborhood of $(\bar{z},\bar{t})$, namely,

$$\bar{y} = h(\bar{z},\bar{t}) \text{ and } g(h(z,t),z,t) = 0 \qquad (5)$$

for all (z,t) in the neighborhood of $(\bar{z},\bar{t})$. Consequently,

we may define, at least locally, a <u>generalized reduced parametric problem</u> as follows: For $t \in T$,

$$\begin{array}{lll} \text{minimize}_z & F(z,t) \triangleq f(h(z,t),z,t) & \\ \text{subject to} & z \geq 0 & \end{array} \qquad (6)$$

where h is the implicit function defined by (5). Notice that the objective function F is twice continuously differentiable in x and t, because both f and h have the same property. It is to be noted that more than one such reduced problem may be defined in the neighborhood of $(\bar{x},\bar{t})$, since a possible choice of basic variables is not necessarily unique.

Now review the second-order sufficiency conditions for isolated local minima of the problems (4) and (6). A statement of the conditions for more general nonlinear programs can be found elsewhere, for example, in [9, p.235], and hence, no proof is given here.

For $t \in T$, if an n-vector x(t) satisfies

$$g(x(t),t) = 0, \qquad a_I \leq x_I \leq b_I$$

and if there exists an m-vector λ such that

$$\nabla_{x_I} f(x(t),t) - \lambda \nabla_{x_I} g(x(t),t) = 0 \qquad a_I \leq x_I \leq b_I$$

$$x_{J_1} = a_{J_1} \qquad \nabla_{x_{J_1}} f(x(t),t) - \lambda \nabla_{x_{J_1}} g(x(t),t) \geq 0$$

and $$x_{J_2} = b_{J_2} \qquad \nabla_{x_{J_2}} f(x(.),t) - \lambda \nabla_{x_{J_2}} g(x(t),t) \leq 0$$

then x(t) is an isolated local minimum of problem (4), where $I \cup J_1 \cup J_2 = \{ 1,2,\ldots, n \}$.

For $t \in T$, if an (n-m)-vector z(t) satisfies

$$\nabla_{z_I} F(z(t),t) = 0, \qquad a_I \leq z_I \leq b_I$$

$$z_{J_1} = a_{J_1} \qquad \nabla_{z_{J_1}} F(z(t),t)\ z(t) \geq 0 \qquad (7)$$

and $$z_{J_2} = b_{J_2} \qquad \nabla_{z_{J_2}} F(z(t),t)\ z(t) \leq 0$$

then $z(t)$ is an isolated local minimum of problem (6), where $I \cup J_1 \cup J_2 = \{1,2,\dots, n-m\}$.

The following two theorems state important relationships between the optimality conditions for problems (4) and (6), on which the algorithms in the next section are based. The proofs can be completed in a manner quite analoguous to that for Theorems 1 and 2 in [11], and hence are omitted.

Theorem 1. For $t \in T$, if $x(t)$ satisfies the second-order sufficiency conditions for a local minimum of problem (4), then there exists a partition $x(t) = [y(t), z(t)]$ such that $a < y < b$ and that $z(t)$ satisfies the second-order sufficiency conditions for a local minimum of the generalized reduced problem (6).

Theorem 2. For $t \in T$ and an $(n-m)$-vector $z(t)$, let h be an implicit function determined by (5) on a neighborhood of $(z(t), t)$. If $z(t)$ satisfies the second-order sufficiency conditions for a local minimum of the generalized reduced problem (6), and if $a \le h(z(t),t) \le b$, then the n-vector $[y(t), z(t)]$ satisfies the second-order sufficiency conditions for a local minimum of the original problem (4), where $y(t) = h(z(t),t)$.

In Theorems 1 and 2, we have tacitly assumed that we know the function h explicitly when handling (7).

In practice, however, we need to solve another nonlinear equation (5) separately to have $h(z,t)$. Motivated by this fact, we consider the following system of equations and inequalities in place of (7) :

$$\begin{aligned} &g(y,z,t) = 0 \\ &H_{z_I}(y,z,t) = 0 \qquad a_I \le z_I \le b_I \end{aligned} \tag{8}$$

$$H_{z_{J_1}}(y,z,t) \geq 0 \qquad z_{J_1}=a_{J_1}$$

$$H_{z_{J_2}}(y,z,t) \leq 0 \qquad z_{J_2}=b_{J_2}$$

where, for the partition $x = (y,z)$, the function $H: R^n \times T \to R^{n-m}$ is defined by

$$H(x,t) = \nabla_z f(x,t) - \nabla_y f(x,t)\ [\nabla_y g(x,t)]^{-1}\ \nabla_z g(x,t).$$

Theorem 3 shows that (7) may be replaced by (8) in the optimality conditions for the reduced problem.

Theorem 3. For $t \in T$, suppose that an n-vector $x(t) = [y(t),z(t)]$ solves (8). If $a \leq y(t) \leq b$ and if

$$v^T \Gamma^T [\nabla^2_{xx} f - \lambda \nabla^2_{xx} g] \Gamma\ v > 0 \tag{9}$$

for all v, such that $v_j=0$ for $j \in \{j\ ;\ H_j(x(t),t) > 0\} \cup \{j\ ;\ H_j(x(t),t) < 0\}$, where $\nabla^2_{xx} f$, $\nabla^2_{xx} g$, λ and Γ are evaluated at $(x(t),t)$, then $x(t)$ satisfies the second-order sufficiency conditions for problem (4). Conversely, if $x(t)$ satisfies the second-order sufficiency conditions for problem (4), there exists a partition of x into y and z such that $a<y(t)<b$ and that $x(t) = [y(t),z(t)]$ satisfies (8) and (9).

Proof.

See [12].

In what follows, we assume the existence and the continuity of $x(t)$ which satisfies the second-order sufficiency conditions for a local minimum of problem (4) for every $t \in T$. Under this assumption, we shall propose two algorithms for solving problem (4) parametrically. Without any loss of generality, let T be the interval $[0,+\infty]$ in the rest of this section.

On the basis of Theorem 1 and 2, we first state the following algorithm which is a straightforward generalization of the Basic Algorithm presented in [11].

Algorithm 1.

Step 1 : Obtain an optimal solution $x(0)$ of problem (4) for $t=0$ by an appropriate method. Choose a sufficiently small number $\gamma \geq 0$ and set $t'=0$. Go to step 2 ;

Step 2 : Partition x into basic y and nonbasic z such that the corresponding basis matrix is nonsingular and all components of $y(t')$ have a value greater than γ. Go to step 3.1 ;

Step 3.1 : Choose sets I and J of indices such that

$$\{j;\ \nabla_{z_j} F(z(t'),t') > 0\} \subset J_1 \subset \{j;\ z_j(t') = a_j\}$$

$$\{j;\ \nabla_{z_j} F(z(t'),t') < 0\} \subset J_2 \subset \{j;\ z_j(t') = b_j\}$$

where $J = J_1 \cup J_2$, and $I = \{1,2,\ldots, n-m\} - J$. Go to step 3.2 ;

Step 3.2 : Obtain the solution $z(t)$ of the system of equations

$$\begin{aligned} &\nabla_{z_I} F(z,t) = 0 \\ &z_{J_1} = a_{J_1} \\ &z_{J_2} = b_{J_2} \end{aligned} \tag{10}$$

where z_I and z_J are the vectors with components z_i, $i \varepsilon I$, and z_j, $j \varepsilon J$, respectively, and solve (5) to get $y(t) = h(z(t),t)$, as t increases from t' to $t = t^\circ \triangleq \min\{t^*,t^{**}\}$, where

$$\begin{aligned} t^* = \sup\{t;\ &a_i + \gamma < y_i(\tau) < b_i - \gamma \ \ \forall i, \text{ for all } \tau, \\ &\text{s.t. } t' < \tau < t < +\infty\} \text{ and} \\ t^{**} = \sup\{t;\ &\nabla_{z_{J_1}} F(z(\tau),\tau) \geq 0,\ \nabla_{z_{J_2}} F(z(\tau),\tau) \leq 0 \text{ and} \\ &a_I \leq z_I \leq b_I \text{ for all } \tau \text{ s.t. } t' \leq \tau \leq t \leq \infty\}. \end{aligned} \tag{11}$$

Go to step 3.3 ;

Step 3.3 : If $t^\circ = \infty$, terminate. Otherwise, setting $t' = t^\circ$, return to step 2 if $t^\circ = t^*$, and return to step 3.1 if $t^\circ = t^{**}$.

Next, we propose another algorithm for finding a parametric solution of problem (1).

<u>Algorithm 2</u>.

<u>Step 1 and 2</u> : Same as those in Algorithm 1 ;

<u>Step 3.1</u> : Choose I and J, such that

$$\{j;\ H_j(x(t'),t') > 0\} \subset J_1 \subset \{j;\ z_j(t') = a_j\}$$

$$\{j;\ H_j(x(t'),t') < 0\} \subset J_2 \subset \{j;\ z_j(t') = b_j\}$$

where $J = J_1 \cup J_2$, and $I = \{1,2,\ldots, n-m\}$.

Go to step 3.2 ;

<u>Step 3.2</u> : Obtain the solution $x(t) = [y(t),z(t)]$ of the system of equations

$$\begin{aligned} &g(y,z,t) = 0 \\ &H_I(y,z,t) = 0 \\ &z_{J_1} = a_{J_1} \\ &z_{J_2} = b_{J_2} \end{aligned} \tag{12}$$

as t increases from t' to $t = t^\circ \triangleq \min\{t^*,t^{**}\}$, where t^* is defined by (11) and

$$t^{**} = \sup\{t;\ H_{J_1}(x(\tau),\tau) \geq 0,\ H_{J_2}(x(\tau),\tau) \leq 0$$

and $a_I \leq z_I(\tau) \leq b_I$ for all τ, such that $t' \leq \tau \leq t \leq +\infty\}$

where $H = [H_I,H_J]$ and

$$H_I = \nabla_{z_I} f - \nabla_y f[\nabla_y g]^{-1}\nabla_{z_I} g = \nabla_{z_I} f - \lambda\nabla_{z_I} g,$$

$$H_J = \nabla_{z_J} f - \nabla_y f[\nabla_y g]^{-1}\nabla_{z_J} g = \nabla_{z_J} f - \lambda\nabla_{z_J} g.$$

Go to step 3.3 ;

<u>Step 3.3</u> : Same as that in Algorithm 1.

Flowcharts of Algorithms 1 and 2 are given in Figures 1 and 2, respectively.

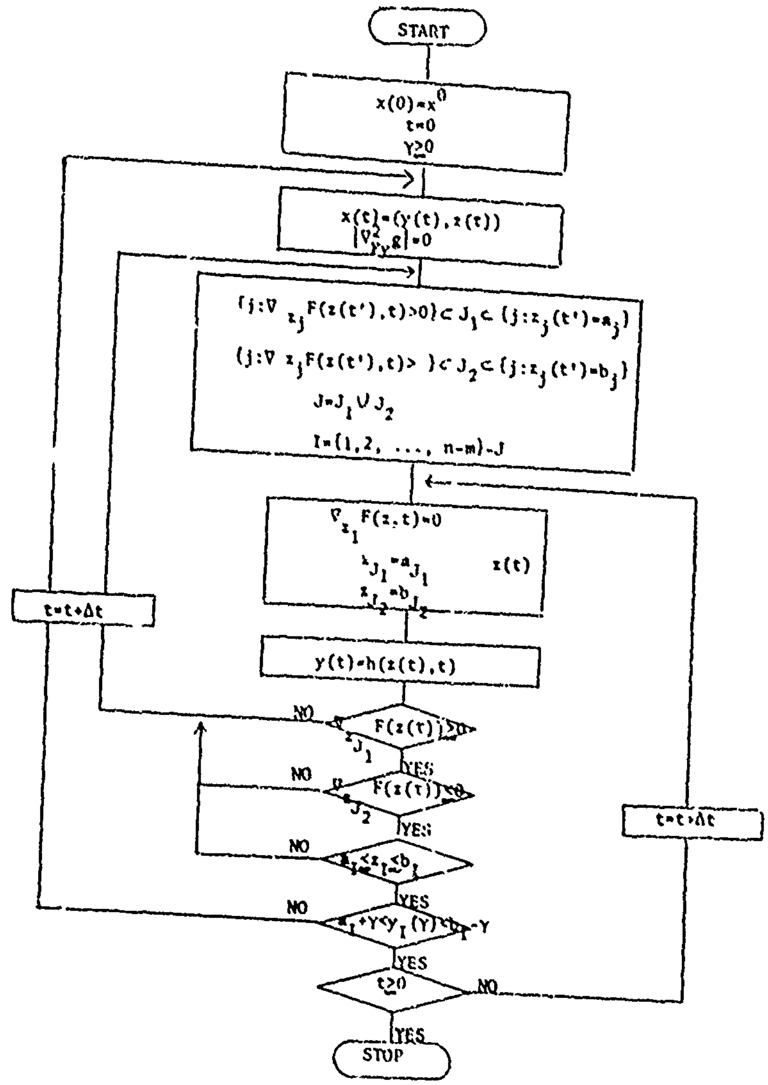


Figure 1. Flowchart of Algorithm 1

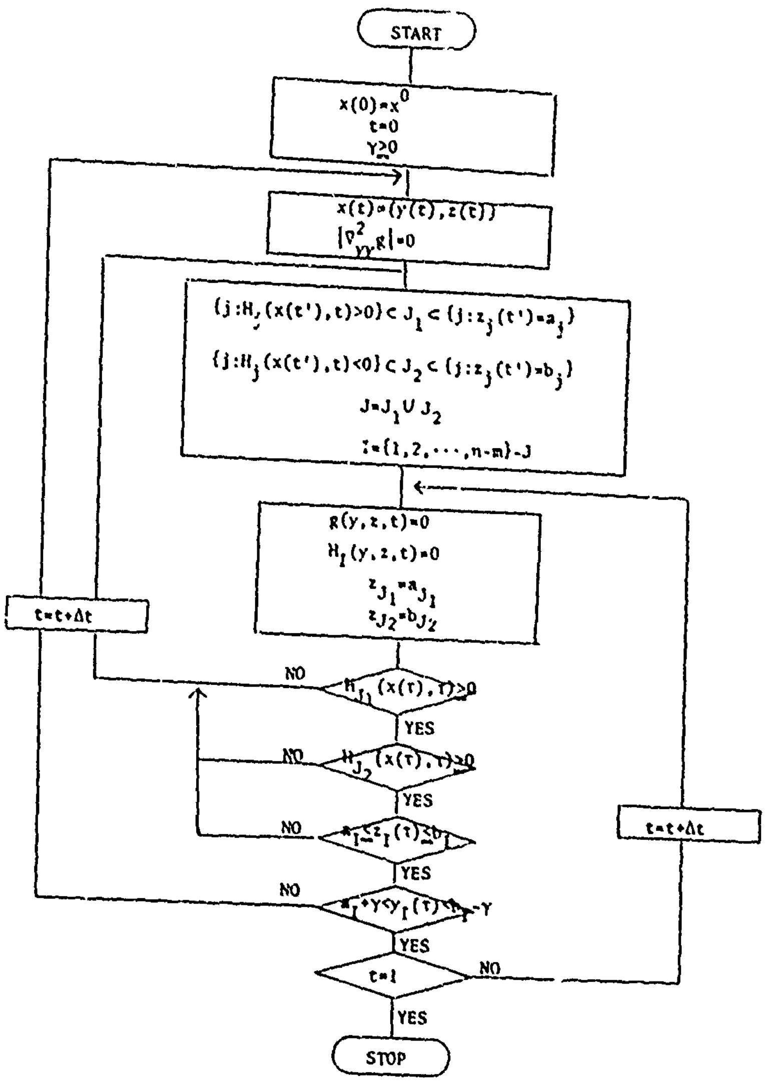


Figure 2. Flowchart of Algorithm 2

It follows immediately from Theorem 3 that Algorithms 1 and 2 are essentially equivalent to each other. To summarize the procedures, we determine parametrically the values of nonbasic variables by solving the equations which are derived from the optimality conditions for the generalized reduced problem. We also monitor the basic variables to decide whether the current basis is changed or not. For more details on the Algorithms the reader may refer to [12].

4. A STRUCTURAL OPTIMIZATION MODEL

Let us suppose that we have a beam with I-shaped cross section as shown in Fig. 3. The maximum stress σ_{max} caused by the bending moment M on the upper and lower surface of this beam is expressed as

$$\sigma_{max} = M \,/\, \left[\frac{T_W \cdot H^3}{12(\frac{H}{2} + T_F)} + (\frac{H}{2} + T_F)B \cdot T_F\right]. \tag{13}$$

In this case, it is an indispensable condition that the stress σ_{max} should not be greater than the allowable stress σ_{all}, i.e.,

$$\sigma_{max} \le \sigma_{all} \tag{14}$$

For the beam with I-shaped cross section, we obtain the following constraints :

$$H - 170T_W \le 0, \tag{15}$$

$$H \le 40, \tag{15'}$$

$$B - 24T_F - T_W \le 0, \tag{16}$$

$$T_W,\ T_F \ge 0.8 \tag{17}$$

where (15) is a constraint on the length and thickness of web to prevent buckling, (15') is an additional constraint due to the capacity of the installation, (16) is a requirement of the effective width of the flange, and (17) is a size limitation caused by the material used. Under the constraints (14) to (17), we have to minimize the cross section area of the beam:

$$A = H \cdot T_W + 2B \cdot T_F \tag{18}$$

For convenience we shall use the following notation :

$$x_1 = T_W,\ x_2 = T_F,\ x_3 = H,\ x_4 = B.$$

Then, we can express the objective function as

$$f = x_1 x_3 + 2x_2 x_4.$$

In addition, by introducing slack variables x_5, x_6, and x_7 we can rewrite the constraint (14) as

$$M(6x_3 + 12x_2)$$
$$- \sigma_{all}(x_1 x_3^3 + 3x_2 x_3^2 x_4 + 12x_2^2 x_3 x_4 + 12x_2^3 x_4) + x_5 = 0,$$

and the constraints (15), (15'), (16), and (17) as

$$x_3 - 170x_1 + x_6 = 0,$$

$$x_4 - 24x_2 - x_1 + x_7 = 0,$$

$$x_1,\ x_2 \geq 0.8,$$

$$x_3 \leq 40,$$

respectively, with x_5, x_6, $x_7 \geq 0$.

In our computational experiment, the value of the bending moment M and the allowable stress σ_{all} are assumed to be 5×10^5 (kg·cm) and 1500 (kg/cm^2), respectively. To apply the deformation method described in the previous section, we regard the above problem as

$$\text{minimize} \qquad f_1(x)$$

$$\text{subject to} \qquad g_1(x) = 0, \quad a \leq x \leq b,$$

and suppose that the initial problem is given as

$$\text{minimize} \qquad f_0(x)$$

$$\text{subject to} \qquad g_0(x) = 0, \quad a \leq x \leq b,$$

where $f_0(x) = ||x - x_0||^2$ and $g_0(x) = g_1(x) + \nu$.

It is noted that the vectors x_0 and ν are determined so as to satisfy the relations $a \leq x_0 \leq b$ and $g_1(x_0) = -\nu$.

Computation tests of the Algorithms in this paper were performed, using double precision, on the FACOM 190 system at the Computation Center of Kyoto University. In our experiment the parameter t was increased until t = 6 for obtaining an optimal solution. The result of the calculation is summarized in the following Table 1.

Table 1. Computation Result Using Algorithm 1

	starting point	optimum point
x_1	2.	0.800000D+00
x_2	2.	0.106995D+01
x_3	30.	0.400000D+02
x_4	20.	0.580354D+01
x_5	0.1941469489	0.0
x_6	3.1	0.960000D+00
x_7	3.	0.206524D+01
f		0.444189D+02

The simplified structural optimization problem has been solved in this section for illustrative purposes only. However, more complicated problems can be solved by this deformation method as well.

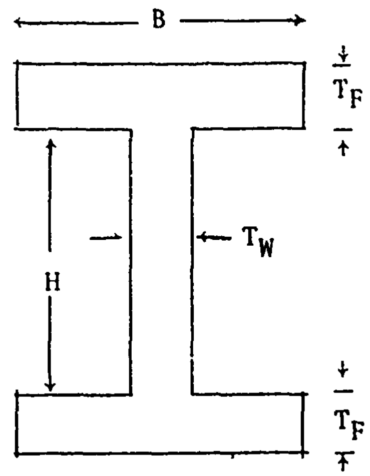


Figure 3. I-shaped Cross Section

5. CONCLUDING REMARKS

In this paper, we have proposed the deformation method for solving nonlinear programming problems and applied it to a structural optimization problem. Finally, we would mention several remarks concerning the method from the viewpoint of application. In the numerical experiments for the structural optimization, the choice of successive parameter values in Algorithms 1 and 2 was important in obtaining the optimum solution. Furthermore, it should always be remarked that the nonsingularity of the Jacobian is an indispensable condition in this deformation method. In addition, it is noted that scaling of the variables of the optimization problems seems to improve the convergence properties of the Algorithms.

REFERENCES

1. J. Abadie and J. Carpentier, "Generalization of the Wolfe Reduced Gradient Method to the Case of Nonlinear Constraints," in Fletcher, R. (ed.), Optimization, Academic Press, New York, 1969, pp. 37-47.
2. W. Dinkelbach, "On Nonlinear Fractional Programming," Management Science, Vol. 13 (1967), pp. 492-498.
3. J. P. Evans and F. J. Gould, "Stability in Nonlinear Programming," Operations Research, Vol. 18 (1970), pp. 107-118.
4. A. M. Geoffrion, "Strictly Concave Parametric Programming, Part I: Basic Theory," Management Science, Vol. 13 (1966), pp. 244-253;"Part II: Additional Theory and Computational Considerations," Management Science, Vol. 13 (1967), pp. 359-370.
5. ________, "Solving Bicriterion Mathematical Programs," Operations Research, Vol. 15 (1967), pp. 39-54.
6. ________, "Proper Efficiency and the Theory of Vector Maximization," Journal of Mathematical Analysis and Applications, Vol. 22 (1968), pp. 618-630.
7. D. J. Hartfiel and G. L. Curry, "Concerning the Solution Vector for Parametric Programming Problems," SIAM Journal on Applied Mathematics, Vol. 24 (1974), pp. 294-296.
8. R. R. Hocking and R. L. Shepard, "Parametric Solution of a Class of Nonconvex Programs," Operations Research, Vol. 19 (1971), pp. 1742-1747.
9. D. G. Luenberger, Introduction to Linear and Nonlinear Programming, Addison Wesley, Reading, Mass., 1973.
10. O. L. Mangasarian, "Nonlinear Programming Problems with Stochastic Objective Functions," Management Science, Vol. 10 (1964), pp. 353-359.

11. H. Mine, M. Fukushima, and Y. J. Ryang, "Parametric Non-linear Programming for Linear Constraints," Kyung-Pan Ron-Chong, Vol. 2 (1977), pp. 135-147.

12. ______________________, "Parametric Non-linear Programming for General Cases and Its Application to Some Problems, "Memoirs of the Faculty of Engineering Kyoto University, Vol. XL, Part 3, (1978), pp. 198-211.

13. S. M. Robinson, "Perturbed Kuhn-Tucker Points and Rates of Convergence for a Class of Nonlinear-Programming Algorithms," Mathematical Programming, Vol. 7 (1974) pp. 1-16.

14. P. Wolfe, "Methods of Nonlinear Programming," in Graves, R. L. and Wolfe, P. (eds.), Recent Advances in Mathematical Programming, McGraw-Hill, New York, 1963, pp. 67-86.

A COROLLARY OF THE LAPLACE TRANSFORM CONVOLUTION THEOREM AND ITS APPLICATION TO ASYMPTOTIC DISTRIBUTIONS

Chang Sup Sung

Software Development Center
Korea Institute of Science & Technology
P.O.BOX 131, Dong Dae Mun
Seoul, Korea

ABSTRACT. It is known that the Laplace-Stieltjes (or just Laplace) transform of a function uniquely determines the limit behavior of the function. The so-called Laplace Transform Convolution Theorem can often be conveniently used to determine the asymptotic limit distributions of convolutions. However, when taking into account a positive constant delivery time τ under the transactions-reporting $\langle Q,r \rangle$ inventory system, the convolution theorem is not satisfactory to get the limit distribution of a lead time demand process $\{D_{(t-\tau,\, t]}\}$, where

$D_{(t-\tau\,, t]} \equiv N_t - N_{t-\tau}$ with the cumulative demand by time t

$\{N_t\}$ which is a discrete-valued continuous-parameter stochastic counting process with sample paths increasing in unit steps. A corollary of the convolution theorem was developed to compute such asymptotic distributions.

1. Introduction

In order to find the asymptotic limit behaviors of some scientific functions, their corresponding Laplace transforms whose definition is given in [3] as follows:

$$L\ \{\varphi(t)\} = \psi(s) = \int_0^{\infty} e^{-st}\ \varphi(t)\ dt, \text{ for } s > 0,$$

where a function φ is of exponential order e^{bt} for constant b, and also sectionally continuous so that it is defined of class $\mathcal{F}$. For example, the limiting behavior of a distribution function F(t) can be found from the equality,

$$\lim_{s \to 0^+} s\ L\{F(t)\} = \lim_{t \to \infty} F(t).$$

Moreover, the so-called Laplace Transform Convolution Theorem in [3] has been conveniently used to determine the asymptotic limit distributions of convolutions. However, in the case of convolutions with a nonnegative parameter $t - \tau$ for $\tau \geq 0$, the theorem itself is not satisfactory to get their limits. Therefore, in this study, a corollary of the convolution theorem is developed for such problems, whose application is made, as an example, to determine the limit distribution of a lead time demand process $D_{(t-\tau,\ t]}$, where $D_{(t-\tau,\ t]} \equiv N_t - N_{t-\tau}$ for the procurement lead time $\tau \geq$ and the cumulative demand by time t $\{N_t\}$. $\{N_t\}$ is assumed to be a discrete-valued continuous-parameter stochastic counting process with sample paths increasing in unit steps.

2. Convolution Theorem

According to [1] and [3], if F(t) is sectionally continuous with at most a finite number of discontinuities and of exponential order e^{bt}, and $F'(t)$ is also sectionally continuous, then

$$L\ \{F'(t)\} \tag{1}$$
$$= s\ L\ \{F(t)\} - F(0^+) - \sum_{i=1}^{n} e^{-st_i}\ [F(t_i^+) - F(t_i^-)]$$

for $s > b$, where $t1, t2, \ldots, t_n$ are the positive abscissas of the points of discontinuity of $F(t)$. In view of (1), following is established that if $F(t)$ is of class $\mathcal{F}$, and further if $F(t)$ has at most a finite number of discontinuities (at $t1, t2, \ldots\ldots, t$) and $F'(t)$ is of class $\mathcal{F}$, then

$$\lim_{s\to o^+} s.\ L\left\{F(t)\right\} = \lim_{t\to\infty} F(t) \quad , \text{ for } b < o , \qquad (2)$$

if either limit exists. Its proof is as follows;

$$\lim_{s\to 0^+} s.\ L\left\{F(t)\right\}$$

$$= \lim_{s\to 0^+} \left[L\left\{F'(t)\right\} + F(0^+) + \sum_{i=1}^{n} e^{-sti}\left\{F(t_i^+) - F(t_i^-)\right\}\right]$$

$$= \int_o^\infty \lim_{s\to o^+} \left[e^{-st} F'(t)\right] dt + F(0_i^+) + \sum_{i=1}^{n} \left\{F(t_i^+) - F(t_i^-)\right\}$$

$$= \int_o^\infty F'(t)\, dt + F(0^+) + \sum_{i=1}^{n} \left\{F(t_i^+) - F(t_i^-)\right\}$$

$$= \lim_{t\to\infty} F(t),$$

since $\displaystyle\int_o^t F'(x)\, dx = F(x)\Big|_o^{t-_1} + F(x)\Big|_{t+_1}^{t-_2} + \ldots + F(x)\Big|_{t+_n}^{t}$

$$= F(t) - F(o^+) - \sum_{i=1}^{n} \left\{F(t_i^+) - F(t_i^-)\right\}$$

and $\displaystyle\int_o^\infty F'(x)\, dx = \lim_{t\to\infty}\int_o^t F'(x)\, dx.$

Eq. (2) shows that if the Laplace-Stieltjes (or just Laplace) transform of a distribution is known, the transform can be directly used to determine the behavior of the distribution function as t tends to infinity.

Using the Laplace transform notation, the so-called Laplace Transform Convolution Theorem shall be stated without proof. Its proof appears in [1], [3] and [6].

<u>Theorem 1</u>

If $F(t)$ and $G(t)$ are of class $\mathcal{F}$, then

$$L\left\{\int_0^t G(t-x)\, F(x)\, dx\right\}$$

$$= L\left\{F(t)\right\} \quad L\left\{G(t)\right\}, \text{ for } \quad s > b,$$

where e^{bt} is the maximum of the exponential orders of $F(t)$ and $G(t)$.

For convolutions with nonnegative parameters $t - \tau$ ($\tau \geqslant 0$) rather than t, the above theorem does not work. Therefore, such problems can be resolved by the next corollary.

<u>Corollary 1</u>

If $F(t)$ and $G(t)$ are of class $\mathcal{F}$, then for $s > b$ and $t \geqslant \tau$

$$L\left\{\int_0^{t-\tau} G(t-x)\, F(x)\, dx\right\}$$

$$= L\left\{F(t)\right\} \quad L\left\{G(t)\right\} - L\left\{(F(t)\right\}\int_0^{\tau} e^{-sy} G(y)\, dy$$

where τ is a nonnegative constant.

Proof : Define

$$I(A) = \iint_A e^{-s(x+y)} \quad F(x)\, G(y)\, dx\, dy$$

$$= \int_0^k e^{-sx} \quad F(x) dx \int_0^k e^{-sy}\, G(y)\, dy$$

such that the region A of integration is illustrated in Fig. 1.

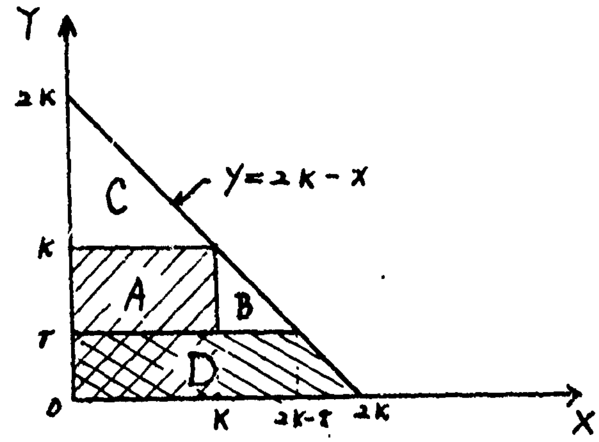


Fig. 1. Illustration of the domain of integration I(A).

Then,

$$L\left\{F(t)\right\} \; L\left\{G(t)\right\}$$

$$= \left(\int_0^{\infty} e^{-st} \; F(t)\,dt\right)\left(\int_0^{\infty} e^{-st} \; G(t)\,dt\right)$$

$$= \lim_{k\to\infty} \int_0^{k} e^{-sx} \; F(x)dx \int_0^{k} e^{-sy} \; G(y)dy$$

$$= \lim_{k\to\infty} I(A)$$

Similarly,

$$L\left\{\int_0^{t-T} G(t-x)F(x)dx\right\} =$$

$$= \lim_{k\to\infty} \int_0^{2k} e^{-st} \left(\int_0^{t-\tau} G(t-x)\, F(x)\, dx \right) dt ,$$

whose integral is equal to a double integral over the triangular region shown in Fig. 2,

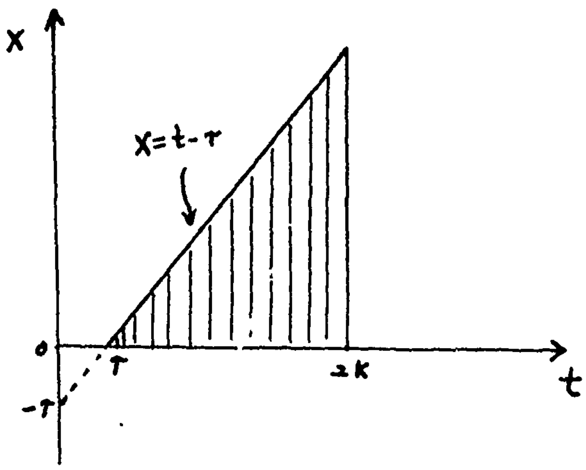


Fig. 2. Domain of integration I(R).

$$= \lim_{k\to\infty} \int_\tau^{2k} e^{-st} \left(\int_0^{t-\tau} G(t-x) F(x) dx \right) dt \text{ , for } t \geqslant \tau$$

$$= \lim_{k\to\infty} \int_0^{2k-\tau} F(x)\, dx \int_{x+\tau}^{2k} e^{-st}\, G(t-x) dt$$

$$= \lim_{k\to\infty} \int_0^{2k-\tau} F(x) dx \int_\tau^{2k-x} e^{-s(x+y)}\, G(y)\, dy,$$

replaced $t-x$ by y,

$$= \lim_{k\to\infty} I(R) \cdot$$

where the region of integration R is composed of the three domains $A \cap \tilde{D}$, B and C ($\tilde{D}$ is the complement of D)

in Fig. 1.

However,

$$L\left\{\int_0^t G(t-x)F(x)dx\right\}$$

$$= \lim_{k\to\infty}\int_0^{2k} e^{-st}\left(\int_0^t G(t-x)F(x)dx\right)dt$$

$$= \lim_{k\to\infty}\int_0^{2k} F(x)dx\int_0^{2k-x} e^{-s(x+y)}\ G(y)\,dy$$

$$= \lim_{k\to\infty} I(R'),$$

where the region of integration R' is composed of the three domains $A\cap\tilde{D}$, B, C and D.

Since $\lim_{k\to\infty} I(R') = L\{F(t)\}\ L\{G(t)\}$,

$$L\left\{\int_0^{t-\tau} G(t-x)F(x)dx\right\}$$

$$= L\{F(t)\}\cdot L\{G(t)\} - \lim_{k\to\infty} I\,(D)$$

$$= L\{F(t)\}\cdot L\{G(t)\}$$

$$-\lim_{k\to\infty}\int_0^{\tau} e^{-sy}\ G(y)dy\int_0^{2k-y} e^{-sx}\ F(x)dx,$$

where given $0 \le y \le \tau$,

$$\lim_{k\to\infty}\left|\int_{2k-\tau}^{2k} e^{-sx}\ F(x)dx\right| \le \lim_{k\to\infty}\int_{2k-\tau}^{2k} e^{-sx}\left|F(x)\right| dx = 0$$

for the convergence of $\int_0^{\infty} e^{-sx} \left| F(x) \right| dx$ for $s > b$,

$$= L\left\{F(t)\right\} \cdot L\left\{G(t)\right\}$$

$$- \lim_{k\to\infty} \int_0^{\tau} e^{-sy} G(y)dy. \left[\int_0^{2k} e^{-sx} F(x)dx - \int_{2k-y}^{2k} e^{-sx} F(x)dx \right]$$

$$= L\left\{F(t)\right\} \cdot L\left\{G(t)\right\}$$

$$- \int_0^{\tau} e^{-sy} G(y)dy \left[\lim_{k\to\infty} \int_0^{2k} e^{-sx} F(x)dx \right]$$

$$= L\left\{F(t)\right\} \cdot L\left\{G(t)\right\} - L\left\{F(t)\right\} \int_0^{\tau} e^{-sy} G(y)dy$$

Thus the proof is complete.

3. Application

The above corollary is applied for finding the limit distribution of one-at-a-time inter-demands under the transactions-reporting $\langle Q, r \rangle$ inventory system discussed in [2]. It is assumed under the inventory system that backorders are allowed, lead time τ is a constant, and customer inter-arrival times with finite mean are independent and identically distributed (iid).

Let the successive inter-arrival times $\{X_i ; i > 1\}$ be defined as $X_1 = t_1$, $X_2 = t_2 - t_1$, ..., $X_n = t_n - t_{n-1}$, ... Then, under the above assumptions, the integer-valued, or counting, process $\{N_t ; t \geq 0\}$ is a renewal counting process generated by $\{X_i\}$. Denote by S_n the renewal epoch of the n^{th} demand (the time of the n^{th} renewal), so that $\{S_n ; n = 0, 1, 2, \dots\}$ are the partial sums of the renewal process $\{X_i\}$, that is,

$$S_n = \sum_{i=1}^{n} X_i \quad , \quad (S_0 \equiv 0).$$

In other words, S_n is the waiting time to the n^{th} demand. Thus, the following relation is obtained;

$$N_t = \operatorname{Sup}\left\{ n \,;\, S_n \leqslant t \right\}$$

and

$$P\left\{ N_t = n \right\} = F_n(t) - F_{n+1}(t),$$

where $F_n(t) = P\left\{ S_n \leqslant t \right\}$ which denotes the n-fold convolution of F with itself so that

$$F_{n+1}(t) = F_n * F(t) = \int_o^t F_n(t-x)\, dF(x)$$

$$= \int_o^t F(t-x)\, dF_n(x) \quad \text{for } n=1,2 \ldots\ldots,$$

$$(F_o(t) \equiv 1 \quad \text{for } t \geqslant o).$$

Denote by μ the mean inter-arrival time $E[X]$.

Now, the next theorem is formed for the limit distribution of $D_{(t-\tau, t]}$ under the $\langle Q,r \rangle$ system, whose proof is done by use of the corollary developed in the previous section.

Theorem 2.

For the continuous-review $\langle Q,r \rangle$ inventory system with backorders allowed, constant lead time $\tau \geqslant 0$, iid customer inter-arrival times with finite mean, and units demanded one at a time,

$$\lim_{t \to \infty} P\left\{ D_{(t-\tau, t]} = k \right\}$$

$$= \begin{cases} \dfrac{\int_o^\tau F_{k-1}(y)dy - 2\int_o^\tau F_k(y)dy + \int_o^\tau F_{k+1}(y)dy}{\mu} & \text{, for } k=1,2.. \\[2ex] 1 - \dfrac{1}{\mu}\displaystyle\int_o^\tau \left[1 - F(x)\right] dx & \text{, for } k=0. \end{cases}$$

Proof : Denote by $Z_{t-\tau}$ the time from $t-\tau$ until the first demand subsequent to $t-\tau$, that is,

$$Z_{t-\tau} = S_{N_{t-\tau}+1} - (t-\tau)$$

Then, under the stated assumptions, following two probability distributions can be found;

$$P\left\{D_{(t-\tau,\,t]} = k\right\}$$

$$= \begin{cases} \int_0^{\tau} P\left\{N_{\tau-z} = k-1\right\} dP\left\{Z_{t-\tau} \leqslant z\right\}, & \text{for } k=1,2,. \\ P\left\{Z_{t-\tau} > \tau\right\}, & \text{for } k=0, \end{cases}$$

and

$$P\left\{Z_{t-\tau} \leqslant z\right\} = F(t-\tau+z)$$

$$- \int_0^{t-\tau} \left[1 - F(t-\tau+z-\xi)\right] dm(\xi),$$

where $m(t) = E\left\{N_t\right\} = \sum_{n=1}^{\infty} F_n(t)$.

Thus, $L\left[P\left\{D_{(t-\tau,\,t]} = k\right\}\right] = L\left[P\left\{N_t - N_{t-\tau} = k\right\}\right]$

$$= L\left[\int_0^{\tau} P\left\{N_{\tau-z} = k-1\right\} dP\left\{Z_{t-\tau} \leqslant z\right\}\right]$$

$$= L\left[\int_0^{\tau}\left\{F_{k-1}(\tau-z) - F_k(\tau-z)\right\}\cdot\right.$$

$$\left.\cdot\left\{f(t-\tau+z) + \sum_{n=1}^{\infty}\int_0^{t-\tau} f(t-\tau+z-\xi)\,dF_n(\xi)\right\}dz\right]$$

$$= L\left[\int_{t-\tau}^{t}\left\{F_{k-1}(t-x) - F_k(t-x)\right\}dF(x)\right.$$

$$+ \sum_{n=1}^{\infty}\int_0^{t-\tau}\left(\int_0^{z=\tau}\left\{F_{k-1}(\tau-z) - F_k(\tau-z)\right\}\cdot\right.$$

$$\left.\cdot\, dF(t-\tau+z-\xi)\right) dF_n(\xi)\Big],$$

replacing $t-\tau+z$ by x,

$$= L\left[\int_{t-\tau}^{t}\left\{F_{k-1}(t-x) - F_k(t-x)\right\} dF(x)\right.$$

$$+ \sum_{n=1}^{\infty} \int_0^{t-\tau}\left(\int_{t-\xi-\tau}^{t-\xi}\left\{F_{k-1}(t-\xi-y)\right.\right.$$

$$\left.\left.\left. - F_k(t-\xi-y)\right\} dF(y)\right) dF_n(\xi)\right],$$

replacing $t-\tau+z-\xi$ by y,

$$= \hat{f}(s)\left\{\int_0^{\tau} e^{-sy} F_{k-1}(y)dy - \int_0^{\tau} e^{-sy} F_k(y)dy\right\}$$

$$+ G(\tau,s)\sum_{n=1}^{\infty}\left(\hat{f}(s)\right)^n ,$$

using Theorem 1 and Corollary 1,

where $\hat{f}(s) = \int_0^{\infty} e^{-st}\, dF(t)$, and

$$G(\tau,s) = \int_0^{\tau} e^{-sy} F_{k+1}(y)dy - \int_0^{\tau} e^{-sy} F_k(y)dy$$

$$+ \hat{f}(s)\int_0^{\tau} e^{-sy} F_{k-1}(y)dy$$

$$- \hat{f}(s)\int_0^{\tau} e^{-sy} F_k(y)dy$$

$$= \hat{f}(s)\left\{\int_0^{\tau} e^{-sy} F_{k-1}(y)dy - \int_0^{\tau} e^{-sy} F_k(y)dy\right\}$$

$$+ G(\tau, s)\frac{\hat{f}(s)}{1-\hat{f}(s)}$$

Therefore,

$$\lim_{t\to\infty} P\left\{D_{(t-\tau,\ t]} = k\right\} = \lim_{s\to 0^+} S \cdot L\left[P\left\{D_{(t-\tau,\ t]} = k\right\}\right]$$

$$= \lim_{s\to 0^+} \frac{s \cdot G(\tau, s) \cdot \hat{f}(s)}{1 - \hat{f}(s)}$$

$$= \lim_{s\to 0^+} \frac{G(\tau, s)\cdot \hat{f}(s) + s\left\{G'(\tau, s) \cdot \hat{f}(s) + G(\tau, s)\ \hat{f}'(s)\right\}}{-\hat{f}'(s)}$$

applying ℓ'Hospital's rule

$$= \lim_{s\to 0^+} \frac{G(\tau, s)\cdot \hat{f}(s)}{-\hat{f}'(s)}$$

Thus, the proof is complete, since

$$\lim_{s\to 0^+} G(\tau, s)$$

$$= \int_0^\tau F_{k-1}(y)dy - 2\int_0^\tau F_k(y)dy + \int_0^\tau F_{k+1}(y)dy,$$

$$\lim_{s\to 0^+} \hat{f}(s) = 1, \text{ and}$$

$$\lim_{s\to 0^+} \hat{f}'(s) = \lim_{s\to 0^+} \int_0^\infty (-x)\ e^{-sx}\ f(x)dx = -E(X).$$

4. Conclusion

As long as the Laplace transform of a convolution is known, its limit behaviors can be verified by use of the Laplace Transform Convolution Theorem. This work shows, however, that when convolutions have nonnegative parameters $t-\tau$ for $\tau \geq 0$, the corollary in section 2 should rather be applied for the corresponding limit behaviors.

The inventory problem treated in this study can also be especially solved by using the so-called Key Renewal Theorem (see [4] and [5]).[1]

Acknowledgement : The author is indebted to professors, H.T. David and B. Vinograde, for helpful conversation in connection with this work.

References

1 Doetsch, Gustav, GUIDE TO THE APPLICATIONS OF LAPLACE TRANSFORMS, D. Van Nostrand Company, New York, 1961.

2 Hadley, G. and T.M. Whitin, ANALYSIS OF INVENTORY SYSTEMS, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

3 Holl, D.L., C.G. Maple and B. Vinograde, INTRODUCTION TO THE LAPLACE TRANSFORM, Appleton-Century-Crofts Inc., New York, 1959.

4 Smith, W.L., RENEWAL THEORY AND ITS RAMIFICATIONS, Journal of the Royal Statistical Society, Series B, Vol 20, PP. 243-302, 1958.

5 Takacs, L., ON A PROBABILITY PROBLEM IN THE THEORY OF COUNTERS, Annals of Mathematical Statistics. Vol 29, PP. 1257-1263, 1958.

6 Widder, D.V., AN INTRODUCTION TO TRANSFORM THEORY, Academic Press, New York, 1971.

1. This work forms part of the author's Ph.D. thesis written at Iowa State University.

LIAPUNOV TECHNIQUE FOR NONLINEAR PROGRAMMING

Prof. P. Venkata Rao+ and Dr. P.A. Janakiraman@

ABSTRACT. This paper presents a new and unconventional approach for locating the equilibrium point of a real-valued nonlinear function of 'n' real variables. Most of the known methods of search depend alrgely on the fact that the stationary points of a function can be found by setting all the first partial derivatives to zero. In the method proposed here, the search is carried out on a positive-definite Liapunov function of the gradients of the objective function with respect to the independent variables. By extending the concept of Asysmptotic Stability of a Sampled Data System, a search method is developed which ensures that the Liapunov function progressively decreases in the direction of movement.

The method described in this paper is conceptually different from the presently known methods and provides an infallible and powerful alternative search technique for the solution of even highly complex Nonlinear Programming Problems.

+Visiting Professor - Department of Electrical Engineering, University of Singapore, SINGAPORE 5

@Lecturer - Department of Electrical Engineering, Indian Institute of Technology, MADRAS - 36 (India)

INTRODUCTION

This paper presents a new approach for locating the equilibrium point of a real-valued nonlinear function of 'n' real variables. In the method proposed here, the search is carried out on a positive definite Liapunov Function of the Gradients of the Objective Function with respect to the independent variables. This method guarantees asymptotic stability of the search movements.

1. THEORY

The general optimization problem including the effect of constraints if any, can be written as:

$$\text{Optimize:} \quad F(x_1, x_2, \ldots\ldots, x_n) \tag{1}$$

In this method, a positive definite Liapunov Function is constructed using the gradients of the Objective Function as the state variables - X of the system under study.

For example: (a) $V = k_1 g_1^2 + k_2 g_2^2 + \ldots\ k_n g_n^2$ (2)

or

$$\text{(b)} \quad V = \sum_{i=1}^{n} k_i . \text{Abs}(g_i) \tag{3}$$

where k_i is +ve.

The new search point with respect to the base point is determined in such a way that the Liapunov Function V progressively decreases and eventually reaches zero, asymptotically. In this new search strategy, divergence of the movements is discouraged.

One can use one-at-a-time Newton Raphson's method for determining the increments for the state variables X so as to minimize the V function. Here, the minimum is identically equal to zero.

$$x_i(\text{new}) = x_i(\text{old}) + \Delta\, x_i \tag{4}$$

$$V = \sum_{i=1}^{n} k_i\, g_i^2 \; : \; k_i > 0$$

$$\delta V/\delta x_i = 2 \sum_{j=1}^{n} k_j \, g_j \, h_{ji} \text{ where } h_{ji} = \delta g_j/\delta x_i \tag{5}$$

$$V(new) = V(old) + (\delta V/\delta x_i) . \Delta x_i \tag{6}$$

Since V is to be reduced to zero,

$$\Delta x_i = -\sigma . \; V/(\delta V/\delta x_i) \tag{7}$$

σ may be assumed to be either 1 or a unimodal search for the values of σ can be carried out so that one gets maximum decrement in V. One can use any other standard search procedure to drive the V function to zero. Incrementation can be either done sequentially or simultaneously for all the state variables x.

2. IMPORTANT FEATURES

1. The V function can be changed after a few iterations.
2. One can switch between two V functions cyclically when steep ridges and valleys are encountered as in Rosenbrock's problem.

3. PROBLEMS WITH CONSTRAINTS

Minimize: $F_{obj}(X)$; $X = (x_i \; ...., \; x_n)$

Subject to: $FL_i \geq 0$; $i = 1, 2, ..., P$

and: $FT_j = 0$; $j = 1, 2, ..., q$

The problem can be cast in Lagrange's form as:

$$\text{Minimize: } P = F_{obj} + \sum_{j=1}^{q} \lambda_j FT_j - \sum_{i=1}^{q} \lambda_i FL_i \tag{8}$$

Subject to the Khun-Tucker Conditions:

$$\lambda_i FL_i = 0; \; \lambda_i \geq 0 \text{ and } FT_j = 0 \tag{9}$$

The vector X must be in the feasible region in the end. A few of the FL_i's may be violated at the start or during the search for Optimum. Such FL's are designated by FV.

The Composite Liapunov Function V may be written as:

$$V = \underbrace{\sum_{r=1}^{n} K_r g_r^{\,2}}_{\text{Gradient Part}} + \underbrace{\left[KL \sum_{i=1}^{p} \lambda_i^2 FL_i^{\,2} + KT \sum_{j=1}^{q} FT_j^{\,2} \right]}_{\text{Kuhn-Tucker Part}} + \underbrace{KV \sum_{m=1} FV_m^{\,2}}_{\text{Violation Part}} \tag{10}$$

Where $g_r = \delta P/\delta x_r$; $KL > 0$; $KT > 0$; $KV > 0$; $K_r > 0$. (11)

The coefficients k_r, KL, KT and KV may be cyclically changed, once in a few iterations (about 2 or 3). Using V and $\delta V/\delta x$, the step size and the direction of movement can be determined. For polynomial type objective function and constraints, short general purpose programme can be written to get the expressions for $\delta V/\delta x$. Further, λ_i and λ_j increases the dimensionality of the problem. One must determine $\delta V/\delta\lambda$ also, to determine the value of incrementing λ's. However, these are easily determined because of the simplicity of the construction of the V function.

4. NUMERICAL EXAMPLE

$$\text{Minimize:} \quad 2x_1 + x_1 x_2 + 3x_2 \tag{12}$$

$$x_1^2 + x_2 - 3 \geq 0 \tag{13}$$

$$x_1 + 2x_2 - 4 \geq 0 \tag{14}$$

$$\text{V Function} = \sum_{i=1}^{2} g_i^{\,2} + \sum_{m=1}^{2} \left[\lambda_m^{\,2} + X\ \mathrm{viol}(m) \right]. FL_m^{\,2} \tag{15}$$

$$X_{viol}(m) = 1 \text{ when } FL_m < 0$$

$$X_{viol}(m) = 0 \text{ when } FL_m \geq 0$$

A problem oriented programme has been prepared and the result obtained with the same is presented in Table 1. To reduce the computation time, the Liapunov Function is modified as:

$$V = K_g \sum_{i=1}^{2} (g_i 2) + KVL \sum_{m=1}^{2} \left[\lambda^2(m) + X_{viol}{}^{(m)} \right]. FL^2(m) \tag{16}$$

The coefficients K_g and KVL are initially chosen in the ratio 1 : A and cycled once in N iterations, i.e., after N iterations K_g : KVL is changed to A : 1. The results are presented in Table 2. It may be noted that a cycling frequency of, once in 2 or 3 iterations gives a fast convergence.

A very large amplitude of vibration (A = 5) makes the convergence rate small whereas a small amplitude of vibrations (A = 1.5) does not cause any appreciable improvement in the convergence rate. For good performance; 'A' in the range of 2 to 3 and N = 2 or 3 is recommended. The above technique can be visualized as the counterpart of signal stabilization method in nonlinear sampled data system for quenching limit cycles.(4).

5. CONCLUSIONS

A new and flexible algorithm has been presented. The motivation for developing this method was provided by the desire to develop a satisfactory algorithm to solve problems cast in the geometric programming-mode. It is hoped that research workers engaged in the field of Optimization will apply this new method to widely different disciplines to test its utility and flexibility.

ACKNOWLEDGEMENT

The Authors wish to place on record their sincere thanks to the Authorities of the Indian Institute of Technology, Madras (India), and the University of Singapore, Singapore for providing the necessary facilities for carrying out the investigations reported in this paper.

REFERENCES

[1] Wilde, D.J., OPTIMUM SEEKING METHODS, Prentice Hall, Englewood Cliffs, New Jersey, 1964.

[2] Himmelblau, D.M., APPLIED NONLINEAR PROGRAMMING, McGraw Hill, 1972.

[3] R.C. Boyer, SINUSOIDAL SIGNAL STABILIZATION, M.S. Thesis, Purdue University, Lafayette, Ind., Jan. 1960.

[4] P.A. Janakiraman, and P. Venkata Rao, ANALYTICAL AND EXPERIMENTAL STUDIES IN STABILIZATION OF DISCRETE SYSTEMS, International Journal of Control, 1970.

The following problems have also been solved using Liapunov Technique.

I. Objective Function: $Y = x_1 + x_2 + 3\,x_1x_2 - x_1^{\,2} - x_2^{\,2}$

Solution: $x_1 = -1.0$; $x_2 = -1.0$; $Y = -1.0$

V Function Selected: $g^2_{\,1} + g_2^2 + g_1\,g_2$

Initial Condition: $x_{10} = 0.3$; $x_{20} = 0.2$

The Value of the V function after one iteration: 3.861
Output at the 20th iteration:

$V = 0.4467.10^{-3}$

$x_1 = -0.9903$; $x_2 = -0.9901$

$y = -0.9999$

CPU Time on IBM 370/155 = 0.15 sec.

II. ROSENBROCK'S PROBLEM NO. 1

(a) Objective Function: $Y = (x_1 - x_2^{\,2})^2 + (1 - x_2)^2$

Solution: The minimum exists at: $x_1 = 1.0$;

$x_2 = 1.0$; $y = 0$

V Function Selected: $g_1^{\,2} + g_2^{\,2} + g_1g_2$

Initial Conditions: $x_{10} = -1.2$; $x_{20} = 1.0$

The value of the V function after 25 iterations = 0.1245
Output after 400 iterations:

$V = 0.4576.10^{-4}$

$x_1 = 0.9976$; $x2 = 0.9921$

$y = 0.1527.10^{-4}$

III. ROSENBROCK'S PROBLEM NO. 2

Objective Function: $Y = 100(x_1 - x_2^2) + (1 - x_2)^2$

Solution: $x_1 = 1.0$; $x_2 = 1.0$; $Y = 0.0$

V Function selected: $A\ g_1^2 + B\ g_2^2 + g_1 g_2$

Initial Conditions: $x_{10} = -1.2$; $x_{20} = +1.0$

A = 0.3 and B=0.9. The values of A and B were interchanged once in 25 iterations.

The value of the V Function for A = 0.3 and B = 0.9 after 25 iterations = $0.1948.10^{+1}$

Output after 3750 iterations:

$$V = 0.1717.10^{-5}$$

$$x_1 = 0.9914;\ x_2 = 0.9957$$

$$Y = 0.1839\ 10^{-4}$$

Table 1

INITIAL CONDITIONS, INFEASIBLE

ITER	X1	X2	g^1	g^2	λ1	λ2	FL1	FL2	F.OBJ	F-Comp =P	V
0	1.0	1.0	0.0	1.0	1.0	1.0	-1.0	-1.0	6.0	8.000	5.0
5	1.108	1.549	-.2641	+.6723E-1	1.034	1.503	-.2232	+0.2061	8.579	8.501	0.2641
10	1.175	1.427	-.1006E-1	+.2273E-1	0.7362	1.708	-.1931	+0.2905E-1	8.307	8.400	0.6058E-1
15	1.256	1.381	-.2895E-1	+.2555E-1	0.643	1.794	-.4055E-1	+0.1798E-1	8.390	8.384	0.4856E-2
20	1.274	1.367	-.8771E-2	+ 5853E-2	0.6065	1.831	-.1150E-1	+0.6758E-2	8.387	8.382	0.4451E-3
25	1.279	1.361	-.1349E-2	+.1663E-2	0.5948	1.841	-.3043E-2	+0.1561-2	8.383	8.382	0.2539E-4
30	1.28	1.36	-.3853E-3	+.6000E-4	0.5924	1.844	-.1186E-2	+0.2918E-3	8.382	8.382	0.2353E-5

Time for 30 iterations = 1.2 secs on IBM 370 Computer

Search: $\Delta x_i = -\sigma(V/\delta V/\delta x_i)$; $\sigma = 1$ (at first),

then σ is so adjusted that $-\Delta V$ is a maximum.

Table 2

AMPLITUDE RATIO 1 : A	FREQUENCY 1 : N	NO. OF ITERATIONS FOR CONVERGENCE
1 : 2	1 : 1	30
1 : 2	1 : 2	16
1 : 2	1 : 3	48
1 : 3	1 : 1	26
1 : 3	1 : 2	14
1 : 3	1 : 3	18
1 : 4	1 : 1	27
1 : 4	1 : 2	32
1 : 4	1 : 3	36
When no cycling was done		25

Error in Prime Variables less than or equal to 0.1

A SIGNAL FLOW GRAPH METHOD OF GOAL PROGRAMMING MODEL

CHUL SOO LEE and GIL CHANG KIM

The Korea Advanced Institute of Science
P.O.Box 150, Chongyangni
Seoul, Korea

ABSTRACT. The rapid expension of computer diffusion, the rapid rate development of software, and the increase of the application of goal programming for decision problem of industrial and economic planning require an efficient algorithm or method which must be systematic and dynamic. Recently many graph theoretic approaches have been developed solving a sparse system. But most of them are based on Gaussian elimination method. Most of the signal flow graph theory was applied for analysis of linear electronic network. This paper presents a signal flow graph method for solving a goal programming model.

1. INTRODUCTION

1.1 Concepts of Goal Programming(GP)

GP is a technique that is capable of handling decision problems that deal with a single goal with multiple subgoals. Often, multiple goals of management are in conflict or achievable only at the expense of other goals. Futhermore, these goals are incommensurable. GP is a linear mathematical model in which the optimum attainment of goals achieved within the given decision environments of the model, namely the choice variable, constraints, and the objective function. Then, a GP problem involving multiple conflicting goals can be formulated as:

$$\begin{aligned} &\text{Minimize} \quad Z = Cd \\ &\text{Subject to} \quad AX + Rd = 0 \\ &\qquad\qquad\quad X,\ d \geq 0 \end{aligned} \tag{1}$$

where A and R are $m \times n$ and $m \times 2m$ matrices respectively;
d is a 2m-component column vector whose elements are deviational variables, d_i^-, d_i^+ such that:
$d = (d_1^-, d_2^-, \ldots\ldots\, d_m^-;\ d_1^+,\ d_2^+, \ldots\ldots\, d_m^+)^T$;
c is a 2m-component row vector whose elements are product of preemptive factors P_j and ∂'s for weighting at the same priority level such that:
$c = (\partial_1 P_{j1},\ \partial_2 P_{j2},\ \ldots\ldots\ldots,\ \partial_{2m} P_{j2m})$. [4]

1.2 Concepts of Signal Flow Graph

A signal flow graph S is a weighted oriented digraph which stands in a one-to-one correspondence with a set of linear equation. It consists of nodes and branches, interconnected at nodes. The weight of a branch b is denoted by f(b). A walk is an alternating sequence of nodes and branches, v_0, b_1, v_1, b_2,..... b_n, v_n in which each branch b_i is $V_{i-1}\ V_i$. A path is a walk in which all nodes are distinct. A loop is a path which has the same first and last nodes. The path gain of a path is the product of the weights of the branches constituting that path. And the loop gain of a loop is the product of the weights of branches constituting that loop. A subgraph is said to be connected if, disregarding the branch directions, it is possible to go from any one node to any other node by only following branches of the subgraphs S_s.

Theorm. For linear system $AX = b$, if the coefficient matrix A of order n is non-singular, then the solution of above linear equation is given by

$$X_i = \frac{\sum_k f(P^k_{(n+1)i}) \left[1 + \sum_{s,t} (-1)^t f(L^k_{st})\right]}{1 + \sum_{w,v} (-1)^v f(L_{wv})} \quad \text{for } i=1,2,\ldots,n \qquad (2)$$

where $P^k_{(n+1)i}$ is the K^{th} path from source node(n+1) to node i in S(Au); L^k_{st} is the S^{th} subgraph of $t(t>0)$ node-disjoint loops which are also node-disjoint with $P^k_{(n+1)i}$ in S(Au); Au is the argumented matrix obtained from A by attaching -b to the right of A; And L_{wv} is the W^{th} subgraph of $V(V>0)$ node-disjoint loops in S(Au). [2]

Above result is a basic solution concept for linear system by signal flow graph. But it is difficult to manipulate the signal flow graph of GP model. A derivation of equation(2) is made by the meaning of the summation of node-disjoint loops and paths as follow:

$$X_i = \sum_{j=1}^{n} I_j T_j \sum_k \Delta_k / \Delta \text{ , for } i = 1,2,\ldots, n \qquad (3)$$

where $\Delta = 1 - \sum_u f(L_u) + \sum_v f(L_v) - \sum_w f(L_w) + \ldots\ldots$

$f(L_u)$ is the loop gain of the u^{th} loop the summation being taken over all loops; $f(L_v)$ is the product of the loop gains of all non-touching loops taken two at a time, and $f(L_w)$ is the product of the loop gains of all non-touching loops taken three at a time, etc [3] ;

I_j is input b_j, $j=1,2,\ldots,n$; T_j is the summation which is taken over all path from source node to node i; $\sum_k \Delta_k$ is the summation of cofactor Δ_k ; Δ_k is obtained by deleting from Δ the loops touched by path $P^k_{(n+1)i}$ for each input i.

2. GRAPH THEORETICAL CONVERSION OF GP MODEL

To solve the GP problem, GP model(1) is reformulated as following ;

Maximize $Z = -Cd$

Subject to $\sum_{i=1}^{m} d_i^- = \sum_{i=1}^{m} 1/r_{ij}(b_i - \sum_{j-1}^{m} a_{ij} X_j +$

$$\sum_{\substack{j=1 \\ j\neq i}}^{m} r_{ij} d_j^- - \sum_{k=1}^{m} r_{ik} d_j^+) \text{ , } X, d \geq 0$$

where $Y = (d_1^-, d_2^-, \ldots, d_m^-)^T$ is a column vector whose elemen· are basic variable. Then, the procedure from the GP to the signal flow graph carry out following step:

1. Each variable in the vector X,d and Y is associated one node.
2. With each constant (bi) is associated a node, a source node.
3. For all i,j if $f(i,j) \neq 0$, there is a branch that orignate at node j and terminate at node i, this branch is assigned a weight $-a_{ij}$ or $-r_{ij}$. If $f(i,j)=0$, no branch is introduced.
4. A branch of unity weight joins the node associated with b_i to the i^{th} basic node, $i=1,2,\ldots\ldots, m$.
5. The objective function goal Z is associated one node. For each deviational variable $d_i(d_i=d_i^-$, $i=1,2\ldots. m$; $d_i{-}d_i^+$, $i= m+1,\ldots., 2m)$, if $\partial_i P_{ji} = 0$, there is a branch that originates at node d_i and terminates at node Z. This branch is assigned a weight- $\partial_i P_{ji}$. If $\partial_i P_{ji}=0$, no branch is introduced.

Above procedure satisfies the relationship of GP model. For example, consider following GP model [4]

$$
\begin{aligned}
\text{Min } Z &= P_1 d_1^- + P_2 d_{11}^+ + 5P_3 d_2^- + 3P_3 d_3^- + P_4 d_1^+ \\
\text{Subject to } \quad & X_1 + X_2 + d_1^+ \qquad + d_1^+ = 80 \\
& X_1 \qquad +d_2^- \qquad = 70 \\
& \qquad X_2 \quad +d_3^- \qquad = 45 \\
& \qquad\qquad d_1 + d_{11}^- - d_{11}^+ \quad = 10 \\
& X_1, X_2, d_1^-, d_2^-, d_3^-, d_{11}^-, d_1^+, d_n^+ \geq 0
\end{aligned}
\tag{4}
$$

Then, corresponding table form and signal flow graph in following;

Table 1. Initial Tableau

	Z			$-P_4$	$-P_2$	$-P_1$	$-5P_3$	$-3P_3$			
	basic	X_1	X_2	d_1^+	d_{11}^+	d_1^-	d_2^-	d_3^-	d_{11}^-	C	F
80	d_1^-	-1	-1	1						80	-1
70	d_2^-	-1								70	-1
45	d_3^-		-1							45	0
10	d_{11}^-			-1	1					10	0
Z_j-C_j (Cm)	P_1	1	1	-1							
	P_2				-1						
	P_3	5	3								
	P_4			-1							

no loop

/ = 1

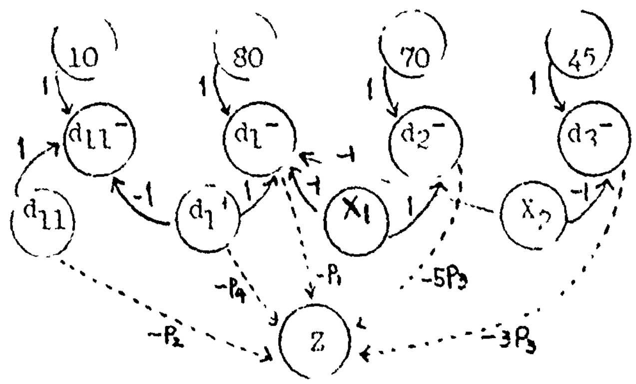


Fig. 1. The signal flow graph of above GP model.

3. METHOD AND EXAMPE.

This method is an iterative method that can obtain the optimal solution of GP model. Therefore, if the basic condition of simplex method is satisfied [5] , optimal solution will be found. The method is divided into several step. First step is to obtain value of basic variables.

Procedure 1. Search all possible loops on the subgraph $S_s(V_B)$[6] and paths from source node to each basic node on $S_s(V_B)$. Calculate Δ , Δ_k and T_j in equation(3). V_B is set of basic nodes

In fig. 1, there is no loop and path between any pair of basic nodes on the subgraph $S_s(V_B)$. Then, Δ is 1. And vector C = $(d_1^-, d_2^-, d_3^-, d_{11}^-)$ = (80, 70, 45, 10). Second step is to determine entering and leaving variables. For determination of entering and leaving variables, it uses the same criterion that the simplex method of GP uses.

Procedure 2. (Determination of entering variable)
For every non-basic node, search all possible path from each non-basic node to the goal node Z and calculate their path gains. If the path pass through the basic node which is contained some loops in procedure 1, delete that loop and path gain from the summation of path gain. Build vector C_m of which component is the summation of above path gains.

Then, there is no loop on the subgraph $S_s(V_B)$. Therefore, each component of vector C_m is each summation of path gains from each non-basic node to the goal node. Hence $C_m = (X_1, X_2, d_1^+, d_{11}^+) = (P_1 + 5P_3,\ P_1 + 3P_3,\ -P_1-P_4,\ -P_2)$ since X_1 is entering variable by entering criterion.[4] For determination of leaving variable, we must search all possible path from node

X_1 to each basic node and calculate their gains. Vector $L = (l_1, l_2, l_3, l_4) = (80, 70, \infty, \infty)$. Therefore, d_2^- is the leaving variable.

Procedure 3. (Determination of leaving variable)
Let X_k be an entering variable. Search all possible path from X_k to each basic node and construct vector F of which component is the summation of above path gains for each basic node. Build vector $L = (l_1 . l_2 . \dots . l_m) = (C_1/-f_1, C_2/-f_2, \dots\dots\dots C_m/-f_m)$. Then, leaving variable is basic variable of corresponding index K which is given by the following formular;

$$\text{index } K = \text{index of } \min_i \left(C_i/-f_i > 0 \right), \quad i=1,2,\dots\dots,m \qquad [5]$$

The next step is to revise tableau by the following procedure.

Procedure 4. (Building of revised tableau)
If there is not a branch from entering variable to leaving variable, it is necessary that a row in the entering variable is added up the row in the leaving variable. Exchange both columns and do not change the weight of branch from entering variable to leaving variable.
The other columns do not change. If the weight of new entering variable is not 1, divide initial constant(rhs) and all other weight in that row by the weight.

Row operation does not influence on linear independency of GP and the determinent of square matrix which consists of basic variables. Therefore, in each step, tableau and signal flow graph are given as follows:

Table 2. 2nd step tableau

	Z	$-5P_3$		$-P_4$	$-P_2$	$-P_1$		$-3P_3$			
IC	basic	d_2^-	X_2	d_1^+	d_{11}^+	d_1^-	X_1	d_3^-	d_{11}^-	C	F
80	d_1^-		-1	1			-1			10	-1
70	X_1	-1								70	0
45	d_3^-		-1							45	-1
10	d_{11}^+			-1	1					10	0
$Z_j - C_j (C_m)$	P_1	-1	1	-1							
	P_2				-1						
	P_3	-5	3								
	P_4			-1							

no loop
$\Delta = 1$

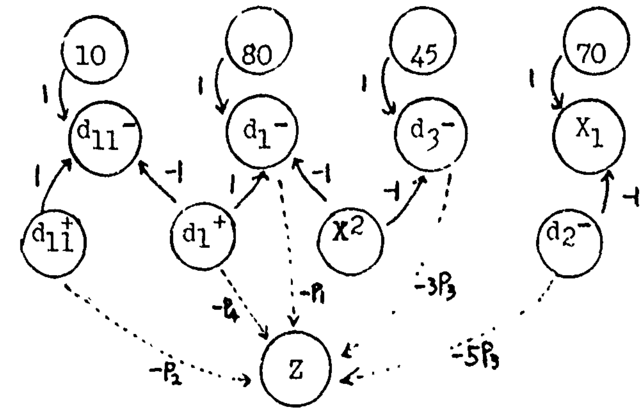


Fig.2. The signal flow graph of table 2.

Table 3. 3rd step tableau

	Z	-5P_3	-P_1	-P_4	-P_2		-3P_3				
IC	basic	d_2^-	d_1^-	d_1^+	d_{11}^+	X_2	X_1	d_3^-	d_{11}^-	C	F
80	X_2		-1	1			-1			10	1
70	X_1	-1								70	0
45	d_3^-					-1				35	0
10	d_{11}^+			-1	1					10	-1
$Z_j-C_j(C_m)$	P_1		-1								
	P_2				-1						
	P_3	-2	-3	3							
	P_4			-1							

no loop

$\Delta = 1$

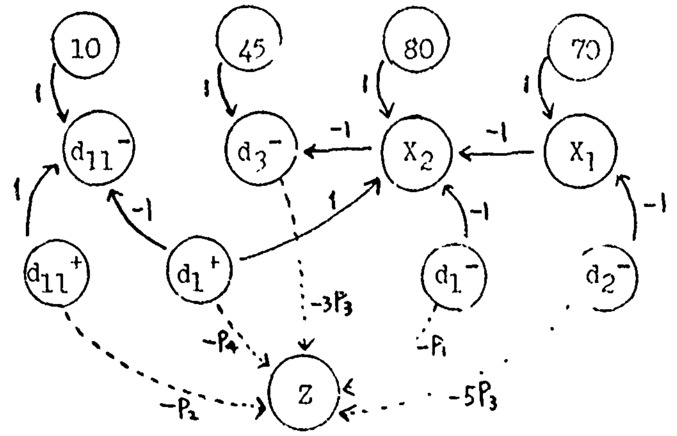


Fig.3. The signal flow graph of table 3.

The last step is to determine whether the solution is optimal. That condition is the following;

> Examine the $C_m(Z_j-C_j)$ coefficient for that row. If there are positive $C_m(Z_j-C_j)$ value in the row, determine whether there are negative $C_m(Z_j- C_j)$ values at a higher priority value for the positive $C_m(Z_j-C_j)$ values in the row of interests, the solution is optimal. [4]

In table 4, $C_m(Z_j-C_j)$ coefficient satisfy above condition. Therefore, the optimum solution is X_1=70, X_2=20, d_1^+=10, d_3^--25.

Table 4. Final tableau

	Z	$-5P_3$	$-P_1$		$-P_2$			$-3P_3$	$-P_4$		
IC	basic	d_2^-	d_1^-	d_{11}^-	d_{11}^+	X_2	X_1	d_3^-	d_1^+	C	F
80	X_2		-1				-1		1	20	
70	X_1	-1								70	
45	d_3^-					-1				25	
10	d_1^+			-1	1					10	
$Z_j-C_j(C_m)$	P_1		-1								
	P_2				-1						
	P_3	-2	-3	-3	3						
	P_4				-1						

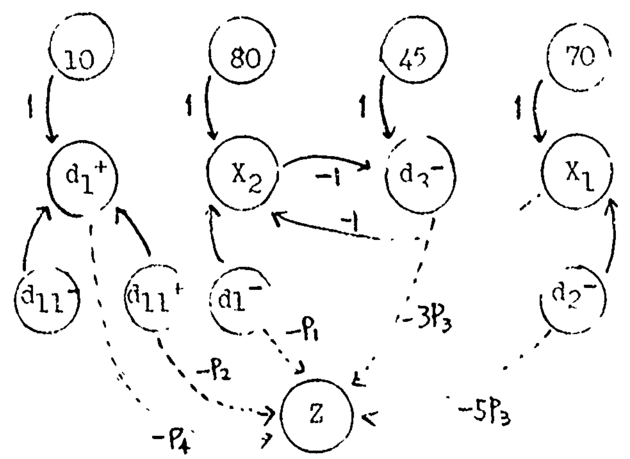


Fig.4. The signal flow graph of table 4.

This method is also useful for post-optimal sensitivity analysis of GP model. Sensitivity analysis in GP is divided into five categories. [4]

Type 1. Change in resources or goal levels(bi)
Type 2. Change in technological coefficient(a_{ij})
Type 3. Addition of a new constraint and goal
Type 4. Addition of a new variable
Type 5. Change in priority variable

Sensitivity analysis for above five categories is very simple in signal flow graph method. Any loop and path do not change in type 1. The product of path gain and the difference(b_i-b_i) will be added up to the value of basic variable of which resource is changed. In type 2, all loops and paths do not change. If the technological coefficient a_{ij} is changed by a'_{ij}, only the gain of paths which are touched with the branch are changed on the subgraph $S_s(V_B)$. If a final node of a path which is through the branch is contained the basic variables, the value of the basic variable is changed. Addition of a new variable or a new constraint may also involve change in the objective function, technological coefficients, and constant of the basic variables. And some new nodes and branches are added to the final signal flow graph. If there is no altered loops, the result is calculated by similar procedures of type 1 or type 2. But if the loops are altered, this case is more complex than other types. In type 5, if all priority factors are changed and constraints are not changed, the vector C_m (Z_j-C_j) is recalculated in final graph. And then the solution is not optimal, more iteration is necessary. Specially, we design an interactive computer based algorithm for sensitivity analysis for GP. Therefore, any type of sensitivity analysis is possible.

4. DISCUSSION

A GP model is a large sparse linear system. This signal flow graph method is more useful for large sparse system. Because large sparse system has many zero coefficients and the zero coefficients are associated zero weighted branches, those branches can be deleted in a signal flow graph. Therefore, this method requires less computer memory than simplex method [4]. This method does not require to build inverse matrix of basic variables. So the number of multiplication and division is less than those of simplex method. Also it is simple to build revised tableau. But this method has a weak point which is the searching of loops and paths in the signal flow graph. But the interconnection of basic variables are not complex. Therefore, the number of loops is limmited due

to the reasous given above. If more efficient algorithm for searching loops and paths is developed, this method is more effi efficient than simplex method in terms of computer memory and excution time.

REFERENCES

1. A.T. Berztiss, DATA STRUCTURES: THEORY AND PRACTICE, Academic Press, 1971.

2. Wai-Kai Chen, APPLIED GRAPH THEORY, North-Holland Publishing Company, 1971.

3. Ernest J. Henley & R.A. Williams, GRAPH THEORY IN MODERN ENGINEERING, Academic Press, 1973.

4. Sang M. Lee, GOAL PROGRAMMING FOR DECISION ANALYSIS, Auberbach Publishers Inc., 1972.

5. Donald M. Simmons, LINEAR PROGRAMMING FOR OPERATIONS RESEARCH, Holden-Day Inc., 1972.

6. Robert Tarjan, "DEPTH-FIRST search and linear graph algorithm". SIAM J. Comput. Vol. 1, No.2, June, 1972.

MANAGEMENT ISSUES AND DECISIONS IN ARMOR COEAS* AND STUDIES

THOMAS B. CAVIN

US ARMY TRADOC SYSTEMS ANALYSIS ACTIVITY
WHITE SANDS MISSILE RANGE
NEW MEXICO

ABSTRACT. Armor System COEAs and studies encompass many activities. During the development of the study plan and execution of the study, a study manager must make many decisions regarding each activity which can impact the study results. This paper identifies the required decisions and discusses the advantages and disadvantages of options. Decisions made in recent armor COEAs and studies are reviewed, and trends in decision making are identified. The decisions selected for discussion include:

1. Determining the scenarios and force levels to be evaluated.

2. Selecting a combat model.

3. Choosing the measures of effectiveness (MOEs).

4. Identifying the systems to be included in the MOEs.

5. Choosing analysis points.

6. Selecting the cost analysis methodology.

7. Determining cost effectiveness relationships.

8. Identifying separate analyses to be conducted.

9. Summing the results and selecting the preferred alternatives.

---

*Cost and Operational Effectiveness Analyses (COEAs).

COEAs and studies reviewed are for tanks, APC's, and AT vehicles. Exact rules or mathematical techniques for making decisions are not included, rather guidelines and considerations are presented. A philosophy or way of thinking is encouraged.

1. INTRODUCTION

1.1. General

Armor System Cost and Operational Effectiveness Analyses (COEAs) and studies encompass many activities. During the development of the study plan and execution of the study, a study manager must make decisions regarding each activity which can impact the study results. This paper identifies the required decisions and discusses the advantages and disadvantages of options. Decisions made in past COEAs are reviewed, and trends in decision making are identified. Throughout this discussion, the term study is used in a broad sense to include both COEAs and studies having the general objective of determining cost and operational effectiveness.

1.2. Background

COEAs consist of operational effectiveness and cost analyses of alternatives to determine the extent they meet the Army's needs. Alternatives are usually rank ordered and a preferred alternative may be selected. The US Army Training and Doctrine Command (TRADOC) has the responsibility for conducting COEAs for the US Army. The TRADOC Systems Analysis Activity (TRASANA) performs most of the analytical work on COEAs for TRADOC. COEAs are performed in accordance with TRADOC requirement documents. These documents provide requirements and gener . information for performing COEAs but it is left to the study manager to provide the detailed on-the-spot direction necessary to conduct the study.

1.3. Study Alternatives

Soon after the need for a COEA has been identified, an initial list of alternatives is provided to TRADOC by Department of the Army. The alternatives may be comprised of equipment systems, organizational schemes, or a combination of these. The final list of alternatives is a result of deliberation of study participants and a Study Advisory Group (SAG) which is convened to advise and guide the study members.

1.4. Management Decisions

The planning and execution of the study methodology may be initiated once the alternatives have been identified.

Listed below are the decisions and choices which must be made by the study manager. Each decision is discussed in the paper.

1. Determining scenarios and force levels to be evaluated.

2. Selecting a combat model.

3. Choosing measures of effectiveness (MOE).

4. Identifying systems to be included in MOE.

5. Choosing an analysis point.

6. Selecting a cost analysis methodology.

7. Determining cost effectiveness relationships.

8. Identifying separate analyses.

9. Summing results and selecting a preferred alternative.

1.5. Recent Armor Studies

During the last four years, six major armor COEAs and studies have been performed. These studies are reviewed in the paper to show the decisions and choices which were made relative to the activities discussed earlier. Figure 1 shows the time frames of these and succeeding studies that are germane to trends discussed in later paragraphs. The analytical agency that performed most of the analyses for each study is identified in Table 1.

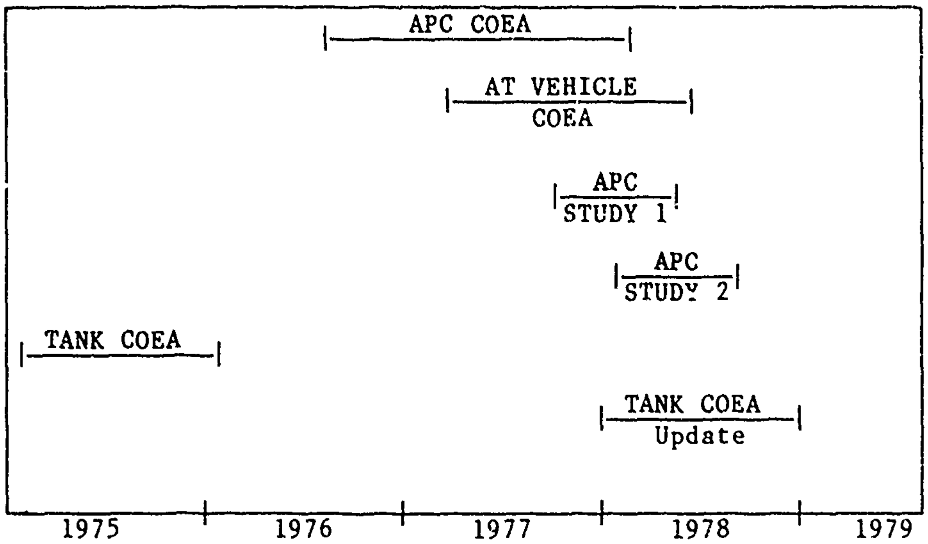


FIGURE 1 RECENT COEAs AND STUDIES

TABLE 1

PRIMARY ANALYTICAL AGENCIES

COEA/STUDY	PRIMARY ANALYTICAL AGENCY
TANK COEA	Concepts Analysis Activity (CAA)
APC COEA	TRASANA
AT VEHICLE COEA	TRASANA
APC STUDY 1	CAA/TRASANA*
APC STUDY 2	TRASANA
TANK COEA Update	TRASANA

*Both agencies listed since both conducted significant portions of the analysis on which the study is based.

2. ORGANIZATION LEVELS AND SCENARIOS

2.1. Introduction

One of the first decisions a study manager makes relative to combat modeling is to determine the set of combat scenarios in which alternative-equipped forces are to be examined. This decision involves organizational levels, types of operations in which the alternatives are to be evaluated, and types of terrain over which the operations are to be conducted. The decision is important because the alternatives must be analyzed across a large enough portion of the spectrum to insure that any performance gaps or shortcomings of the alternative are not overlooked.

2.2. Background

With the currently available combat models, the lower the level of the organization being simulated, i.e., company and battalion, the more detailed and higher fidelity the models. The division-level models may use the results of high-resolution lower-level models for determining battle outcomes in the division game. The high-resolution lower-level model outcomes are adjusted for the different force compositions and force ratios needed for the division game. This is done by interpolating from the results of a library of runs conducted with the high-resolution lower-level force models or by re-allocating the firepower within a single scenario to generate the outcome for situations needed. TRASANA is currently using the latter technique with a re-allocation program titled COMANEX. Consequently, a higher-level model is somewhat less exact in results than a high-resolution lower-level model since the tactics are portrayed more realistically for the exact situation in the high-resolution scenarios and would not be entirely applicable for the interpolated or re-allocated situations. High-resolution lower-level simulation must be available before a division evaluation can be performed with the Division Battle Model (DBM).

2.3. Division Analysis Considerations

A division-level analysis does provide results, although with less resolution, over a range of attacker/defender force ratios and force mixes (tank heavy, mechanized-infantry heavy, and balanced). This, to a certain extent,

tends to weight the scenarios as they might occur in a division battle. Another advantage of a division evaluation is the ability to evaluate the mobility of a system moving, as part of a force, to a critical blocking, reserve, or alternate battle position. Also, differences in effectiveness of alternative equipped forces observed at lower organizational levels may not be observed in a division level evaluation, since differences in capabilities of alternatives may be compensated for by the skillful employment of forces. An example of this might be the application of reserve strength to areas that might be weakened when a less effective alternative is present in the force.

2.4. Scenario Considerations

The decision of which scenarios to use for the various force organizational levels must also be made. Choices include offensive/defensive scenarios and the types of terrain for each. If only defensive or offensive scenarios were evaluated there is a possibility of recommending a weapon system on the basis of its performance in only that one role. This would not be appropriate if in actual use the system is required to conduct both offensive and defensive operations. The same is true of terrain. Care must be taken to preclude the preference of a system evaluated in a single type of terrain when operationally it must be used in various types of terrain.

2.5. Combat Models and Terrain

TRADOC now has a selection of simulations and wargames for different organizational levels. Also, a large amount of different types of terrain has recently been digitized. Consequently, from early 1978 on these can no longer be considered limiting factors. Computer models and wargames will be discussed in additional detail in the following chapter.

2.6. Review of Scenarios Used In Recent Armor Studies

The organizational levels and scenarios that were used for analyses in recent studies are presented in Table 2.

TABLE 2

SELECTED FORCE LEVEL SCENARIOS

COEA/STUDY	CO TM/BN TF DEF	CO TM/BN TF CNTR-ATK	CO TM/BN TF ATK	CAVALRY/SCOUT MVMT TO CON	CAVALRY/SCOUT COVER FORCE	DIV	THEATER/ CORPS
TANK COEA	X		X			X	X
APC COEA	X	X	X	X	X	X	
AT VEHICLE COEA	X		X		X		
APC STUDY 1	X	X	X		X		
APC STUDY 2	X	X	X		X	X	
TANK COEA UPDATE	X	X					

An examination of the table indicates that study managers are conducting both attack and defend scenarios at the CO TM/BN TF level with high resolution models. Also, the trend in more recent studies is to conduct a division level battle to determine if the effectiveness differences observed at lower organizational levels are also present at the division level where various force ratios and compositions are utilized and where there is more freedom for the commander to compensate for the various alternatives. There is also a trend not to rely on one type of terrain but rather to evaluate the alternative equipped forces in different types of terrain.

3. SELECTING COMBAT MODELS/WARGAMES

3.1. Introduction

The analysis tools for performing the force-on-force combat modeling must be chosen after the organizational levels and tactical scenarios have been selected. The choice

of the appropriate analysis tool is important since it must be able to faithfully represent the performance differences of the systems being evaluated. The combat simulations and wargames currently available in the TRADOC inventory and the organization levels for which they are used are listed in Table 3.

TABLE 3

MODELS/WARGAMES AVAILABLE BY LEVEL OF EVALUATION

ORGANIZATION LEVEL	COMBAT SIMULATION	WARGAME
Company TM/Battalion TF	CARMONETTE	BATTLE
	DYNTACS	CIB
	TRACOM	
Division		DBM*
Theater	CEM**	

*Uses CARMONETTE results for individual battle outcomes within the division battle.

**Not in TRADOC Inventory; Resides at CAA.

3.2. Model and Wargame Descriptions

A brief description of each model and wargame is presented below:

3.2.1. CARMONETTE — A Monte Carlo, event-sequenced, battalion-level armored combat model which can simulate a realistic engagement between two combined arms forces for up to one hour of battle time.

3.2.2. Dynamic Tactical Simulator (DYNTACS-X) — A two-sided, small-unit, high-resolution, dynamic, Monte Carlo, event-sequenced, highly interactive, land combat simulation. The model can represent battalion- or smaller-size armor and mechanized units.

3.2.3. TRACOM — A deterministic (expected value) type battle model designed for simulating force-on-force tank/antitank engagements between battalion-size units.

3.2.4. Battalion Analyzer and Tactical Trainer for Local Engagements (BATTLE) — An open, two-sided, computer-assisted, Monte Carlo, manual wargame played on a three-dimensional terrain board. The wargame can simulate situations involving two opposing tanks up to a battalion task force (including dismounted infantry) opposed by a division.

3.2.5. Close-in Battle (CIB) — An open, two-sided, computer-assisted manual wargame. It is played on a two-dimensional, flattened, hexagonal-grid game board and has resolution down to the fire team level.

3.2.6. Division Battle Model (DBM) — A manually operated, computer-assisted wargame designed to evaluate a division-size unit against appropriate opponents in varying tactical situations. The model provides results of combat interactions involving company-size ground elements, supporting fires, and individual aircraft.

3.2.7. Concepts Evaluation Model (CEM) — A fully automated deterministic computer simulation of theater-level non-nuclear warfare. The model was developed as an analytical aid for addressing theater force structure problems.

3.3. Scenario Setup and Production Times

Nominal times for setting up the scenarios and conducting the simulations and wargames for the analytical tools of the TRADOC community are shown in Table 4.

TABLE 4

SCENARIO DEVELOPMENT AND PLAY TIME

SIMULATION	DEVELOP/CHECKOUT SCENARIOS	CONDUCT SIMULATIONS (COMPUTER TIME)
CARMONETTE	6-9 months	3-6 hours[1]
DYNTACS	6-9 months	20 hours[1]
TRACOM[2]	6 months	40 minutes
WARGAME	**TRAIN PLAYERS AND CONTROLLERS**	**CONDUCT WARGAMES**
BATTLE	1-3 weeks	4-5 weeks[3]
CIB	1 week	6-8 weeks[3]
DBM	1 week	10-12 weeks[3]

[1]Twenty replications.

[2]TRACOM is only run for one replication since it is a deterministic model.

[3]Six replications.

3.4. Simulation and Wargame Considerations

Some of the required choices are whether to use simulations or wargames to perform the analyses and, if using simulations, whether to use deterministic or stochastic models. The wargames have the advantage of more freedom of maneuver, good visibility into the game itself, and little required scenario preparation time. The disadvantages are that they require tying up a group of personnel over a somewhat lengthy period of time. This time is required to train the players and control personnel and conduct the necessary number of replications of the battle required for statistical analyses. Furthermore, a learning process can occasionally be introduced into the results of the game if the same players are used for each replication; therefore, unless adequate precautions are taken, the replications would not be

independent trials. There also appears to be a tendency among some decision makers to lend more credibility to simulations than to wargames because there is less player influence and subjectivity in the simulations. The point is debatable, of course, but there is some agreement that the replications of the simulations are independent trials, and the alternatives are usually given more equal treatment in simulations than in games.

3.5. Stochastic and Deterministic Model Considerations

If the decision is made to conduct computer simulations, the choice must be made between the use of deterministic or stochastic models. The deterministic models generally run faster since additional replications are not required; however, since they are usually based on differential equations of combat, they are more abstract than the stochastic models. In stochastic models, each event such as detection, firing, and killing of individual systems is a discrete event. Each event is available and may be examined individually by an analyst; consequently the stochastic models provide greater visibility into the conduct of the battle. Also, because of the discrete feature of the events, the cause and effect relationships are easier to understand and the results are easier to interpret. The stochastic models, however, run longer and thus require more computer time since a number of replications are required to conduct statistical analyses. In addition to those previously mentioned, the advantage of computer simulations is that once the battle is set up on the machine, any number of replications can be conducted (the deterministic simulations only require one run per alternative). A disadvantage is that the simulations require considerable time to prepare the scenarios and adjust them until the desired battle tactics are portrayed. Generally, tactics may be represented more exactly in a wargame than in a simulation. Also, the simulations, comprised of digital code, are difficult to modify and debug when modifications are necessary and are "down" when debugging is required. They also must compete for computer resources with an Activity's other studies.

3.6. DBM — CARMONETTE Interface

As stated previously, running the DBM wargame requires the results of a high resolution simulation to represent the individual unit battles within the division battle. It is currently programmed to use the output from CARMONETTE;

therefore, if a decision is made to use DBM for the division battle, the CARMONETTE results from a previous study must be available and applicable for adaptation to the study or a library of CARMONETTE battles must be generated specifically for the study.

3.7 Additional Considerations

There are also other considerations in the selection of a simulation which have to do with the fidelity of various features within the model and which depend on the alternatives being examined. For example, if there is a significant difference in mobility between alternatives, then a model should be used which can faithfully reproduce the mobility differences. Other alternative features that must be considered relative to model capabilities are lethality, vulnerability, exposed area, command and control, etc.

3.8 Review of Models Used in Recent Armor Studies

The major models and wargames that were used in the studies being reviewed are shown in Table 5. An examination of the table shows that most of the recent studies have depended on simulations at the Co team/Bn TF level and some also include a division battle. The wargames, when used, were for supplementing the high resolution simulations which, in many cases, could not examine the specific feature of interest of the alternative. They were also used in some instances where sufficient time was not available to develop and check out a scenario, and the wargames were run in parallel to meet a deadline. The simulations were usually considered to be the main analytical tools. Relative to the high resolution models, the trend is for study managers to lean more toward the more transparent stochastic models and away from the more abstract deterministic models.

TABLE 5

SELECTED MODELS/WARGAMES

STUDY	COMPANY TEAM/BATTALION TASK FORCE					DIV	THEATER
	CARMONETTE	DYNTACS	TRACOM	BATTLE	CIB	DBM	CEM
TANK COEA	X	X					X
APC COEA	X		X			X	
AT VEH. COEA			X				
APC STUDY 1	X		X				
APC STUDY 2	X	X		X	X	X	
TANK COEA Update	X						

4. CHOICE OF MEASURES OF EFFECTIVENESS

4.1. General

Measures of effectiveness (MOEs) are used for comparing the performance of one alternative with that of another. The choice and proper use of MOEs are important since they form the basis for the conclusions and recommendations that are presented to the decision makers. The choice of an inappropriate MOE can invalidate the comparisons of performance between alternatives, which in turn can lead to misinterpretation of results and to erroneous decisions. Some of the MOEs commonly used in armor studies are defined below. Battle trend is also important since it provides an indication of which side is considered to be winning the battle.

Loss Exchange Ratio (LER) =

$$\frac{\text{ORANGE SYSTEM LOSSES}}{\text{BLUE SYSTEM LOSSES}} \quad (1)$$

Fractional Exchange Ratio (FER) =

$$\frac{\text{PERCENT ORANGE SYSTEM LOSSES}}{\text{PERCENT BLUE SYSTEM LOSSES}} \tag{2}$$

$$= \frac{\dfrac{\text{ORANGE SYSTEM LOSSES}}{\text{TOTAL INITIAL ORANGE SYSTEMS}}}{\dfrac{\text{BLUE SYSTEM LOSSES}}{\text{TOTAL INITIAL BLUE SYSTEMS}}} \tag{3}$$

$$= \frac{\text{LER}}{\text{INITIAL FORCE RATIO}} \tag{4}$$

Specific Exchange Ratio (SER) =

$$\frac{\text{TOTAL ORANGE SYSTEMS KILLED BY BLUE SYSTEM } i}{\text{TOTAL BLUE SYSTEM } i \text{ KILLED BY ANY ORANGE SYSTEM}} \tag{5}$$

System Contribution (SC) =

$$\frac{\text{ORANGE SYSTEMS KILLED BY BLUE SYSTEM } i}{\text{TOTAL ORANGE SYSTEM LOSSES}} \tag{6}$$

The LER and FER are force-type MOEs which describe how well the forces performed when equipped with the study alternatives in a combined arms scenario. The SER and SC are measures of system effectiveness which measure how well that type of system performed in the same environment.

4.2. Review of MOEs Used in Recent Armor Studies

A compilation of the MOEs that were used in each study identified in this paper is shown in Table 6.

TABLE 6

MOEs FOR SELECTED STUDIES

COEA/STUDY	LER	FER	SER	SC
TANK COEA	X		X	
APC COEA	X	X		
AT VEH. COEA	X	X	X	
APC STUDY 1	X	X		X
APC STUDY 2	X	X	X	X
TANK COEA Update	X	X	X	X

Most study managers agree that force MOEs are the more important since they describe how effectively a force performed when it was equipped with the various alternatives. Systems fight as members of a force and how they enhance or detract from force effectiveness is the crux of the matter, not how the performance of one system compares with that of another when considered in isolation. This is the very reason why the effectiveness models being used today play two opposing combined arms forces.

4.3. Force Level MOEs

Study managers currently prefer the FER to the LER (it is simply the LER multiplied by a constant multiplier, the inverse of the initial force ratio). The FER can be more directly interpreted by decision makers since the force can generally be interpreted as winning the battle if FER > 1 and normally losing the battle if FER < 1. This is not as apparent when using the LER. For example, a defensive force achieving a favorable LER of 4:1 would be considered to be doing well; however, if the offensive force began the assault with an initial force ratio of 6:1, the battle trend would still favor the offense.

4.4. System Level MOEs

The SER and SC are primarily system-type MOEs that are used to gain insights regarding the system's performance within the force. Occasionally, there is a tendency among

study members to use system level MOEs when they cannot discriminate between alternative systems with force level MOEs. This tendency can lead to erroneous conclusions, as evidenced by the following example: A situation might exist where the SC is 3:1 for Alternative 1 and 4.5:1 for Alternative 2. If the SC were used as the primary study MOE, System 2 would be considered 50 percent more effective than System 1. The FER for the two forces containing the systems, however, might be equal. Therefore, the force was not actually enhanced by System 2.

4.5. Other MOEs

Many other MOEs are available and are occasionally advocated by study members. These might be ammunition expenditure, rounds fired, casualties, or movement rates. All of these results are useful and their study will normally aid in explaining study results and cause and effect relationships. Most study managers agree, however, that this is where their utility lies and not in comparing the effectiveness of one system with that of another. This comparison should be made with force-type MOEs.

5. SYSTEMS INCLUDED IN MEASURES OF EFFECTIVENESS

5.1. Introduction

Once the study MOEs have been selected, the decision must be made as to which systems will be included in the calculations. This decision is important since the inclusion or omission of various systems may result in different MOE values.

5.2. Calculations of Measures of Effectiveness

The MOE values are obtained using Equations (1), (2), (5), and (6) defined in paragraph 4.1. and using the data from the killer-victim (K-V) scoreboards. These scoreboards are developed using data from a combat model. An example of the K-V scoreboards is shown in Table 7.

TABLE 7

SAMPLE K-V SCOREBOARD

BLUE (DEFENSE) KILLER	ORANGE (OFFENSE) VICTIM						
	TANK	DEDICATED AT SYS	APC	MAW	HELI-COPTER	LAW	TOTAL
Tank	15	3	5				23
Dedicated AT Sys	5	2	3				10
APC	10	2	10				22
MAW	3		2				5
Helicopter	10	3	5				18
LAW							
Mines	3		2				5
Artillery	10	4	5				19
AD Weapons					6		6
Total	56	14	32		6		108

ORANGE (OFFENSE) KILLER	BLUE (DEFENSE) VICTIM					
	TANK	DEDICATED AT SYS	APC	MAW	HELI-COPTER	TOTAL
Tank	3	1	2	2		8
Dedicated AT Sys	1		1			2
APC	1	2	1	2		6
Helicopter	2		2			4
Artillery	1	1	2	1		5
AD Weapon					1	1
Total	8	4	8	5	1	26

The K-V scoreboard tables contain the victim counts to be used in the MOEs. The controversy regarding which systems to count in MOEs is not which killers to count but which systems to count as victims. All systems are considered killers, but only the victims are summed and used in the MOE

calculations. It is the study manager's decision as to which systems to use for the calculations.

5.3. Initial Force Ratio

The same systems used in the MOE calculations are used for determining the initial force ratio before the battle starts. Examples of an initial force ratio tabulation and of a set of MOE calculations are shown in Figures 2 and 3. The data used in the calculations are from the Table 7 K-V scoreboards. The initial force ratio discussed earlier is the constant that relates FER to LER in Equation (4) of paragraph 4.1.

	BLUE		ORANGE
12	Tanks	100	Tanks
15	APC	50	APC
10	Dedicated AT	30	Dedicated AT
14	MAW		
51	Total	180	Total

$$\text{Initial Force Ratio} = \frac{180}{51} = 3.53$$

FIGURE 2 INITIAL FORCE RATIO

$$\text{LER} = \frac{108}{26} = 4.15$$

$$\text{FER} = \frac{60\%}{51\%} = 1.18$$

$$\text{SER (Blue Tanks)} = \frac{23}{8} = 2.88$$

$$\text{SC (Blue Tanks)} = \frac{23}{108} = 21\%$$

FIGURE 3 SAMPLE MOE CALCULATIONS

The numerator and denominator of the initial force ratio represent combat power before the battle starts. The losses in the FERs represent losses of combat power suffered by the adversaries as a result of the battle.

5.4. Review of Systems Used in Recent Armor Studies

The systems counted in the MOEs and initial force ratios for the studies addressed in this paper are shown in Table 8.

TABLE 8

SELECTED SYSTEMS FOR MOE

	TANKS	DEDICATED ATGM	IFV	MAW	LAW	AAH	OTHER
TANK COEA	X	X	X	X	X	**	
APC COEA	X	X	X		X*	**	
AT VEH. COEA	X	X	X	X		**	
APC STUDY 1	X	X	X	X		**	
APC STUDY 2	X	X	X	X		X	X
TANK COEA UPDATE	X	X	X	X		X	X

*Short range dismounted scenario only.

**Not played in scenario.

Most study managers agree that, in armor studies, major direct-fire tank and antitank systems are the systems that should be used in MOEs. There are advocates, however, for including medium-range antitank weapons (MAWs) and these have been included in some recent armor studies. When analyzing short-range scenarios that are characterized by close terrain and dismounted infantry combat, there are advocates for including short-range antitank weapons in the MOE calculations. They were used in one TANK COEA and in one scenario of the APC COEA. Advocates for including medium- and short-range weapons state that the systems can and do kill at the scenario battle ranges and in these scenarios they do achieve some kills. The author proposes that a broader interpretation of loss of combat power be taken beyond the immediate battle being analyzed. Combat power is then the ability to fight not only in a specific kind of battle in a specific type of terrain but in prior and succeeding battles, both defensive and offensive in nature, and in different types of terrain to include close, moderate, and open. It is further proposed that this interpretation be used for all scenarios in the study. The

systems that would be counted in armor studies then would be direct-fire tank and antitank systems which are effective at long and short ranges in defensive and offensive operations.

5.5. Summary

The choice of which systems to include in MOE calculations lies initially with the study manager and finally with the reviewers and decision makers using the study results. If there are strong arguments for including or excluding a system in the MOEs, then the study manager is urged to calculate the results with and without that system so that its impact on the study results is available and open to review.

6. CHOICE OF AN ANALYSIS POINT

6.1. General

An analysis point must be selected after the combat simulations or wargames are completed and after the tacticians and analysts are both satisfied with the correctness of tactics, equal treatment of alternatives, and fidelity of the battle. This is the point at which the K-V scoreboards are constructed, system losses are totaled, and MOE calculations are made. Some of the choices for selecting analysis points are:

1. Equal time.

2. Equal ORANGE or BLUE losses (%).

3. End of game.

4. Tactical considerations.

(i) Mission completion.

(ii) Mission abort.

None of the above listed analysis points is applicable for every situation but any may be correct for any given specific situation.

6.2. Time Plot

A time plot of the MOE is often a useful tool since it provides a visual indication of what is happening in the

battle for each of the various alternatives, both singly and in unison. An examination of the plot can keep the study manager from selecting a non-appropriate analysis point which would lead to incorrect comparisons of alternative results. A sample time plot for a single alternative, presented in Figure 4, is annotated at significant events to describe what is happening in the simulated battle at that time.

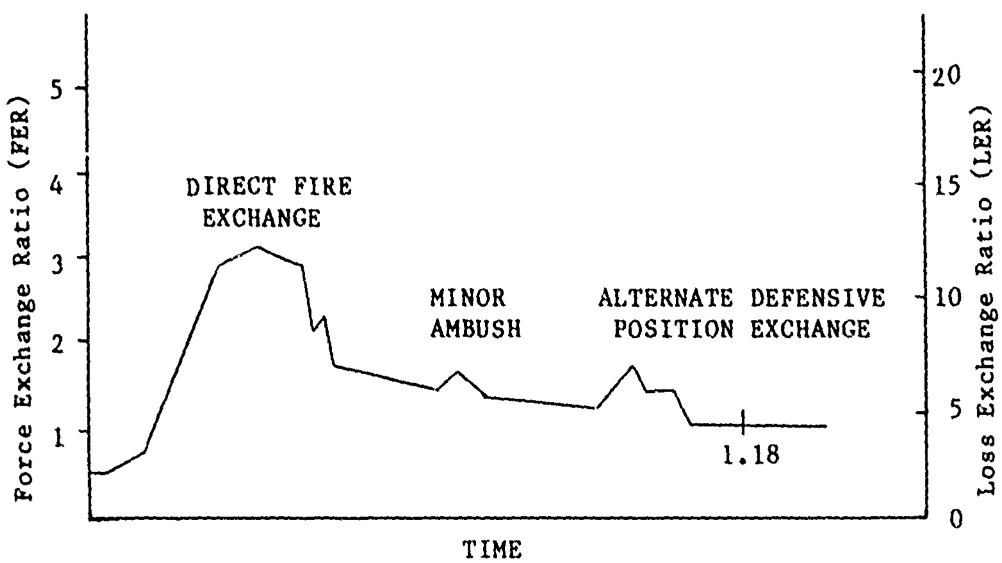


FIGURE 4 ANNOTATED MOE PLOT

6.3 Analysis Points

Time plots of several alternatives for the same battle are presented in Figures 5 and 6. A study of alternatives with equal mobility and different effectiveness is presented in Figure 5, and study results where vehicle alternatives have different mobility and effectiveness are shown in Figure 6.

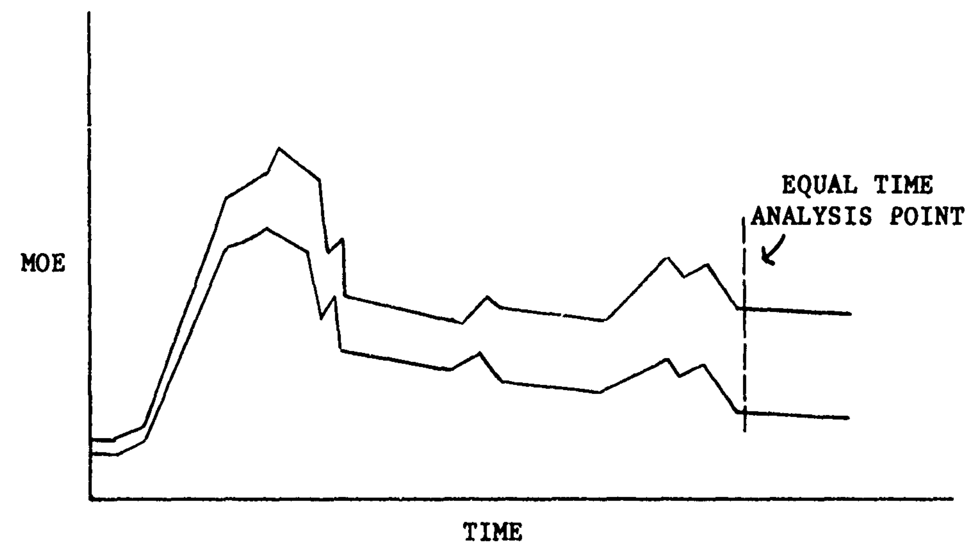


FIGURE 5 MOE PLOT OF EQUAL MOBILITY, DIFFERENT EFFECTIVENESS ALTERNATIVES

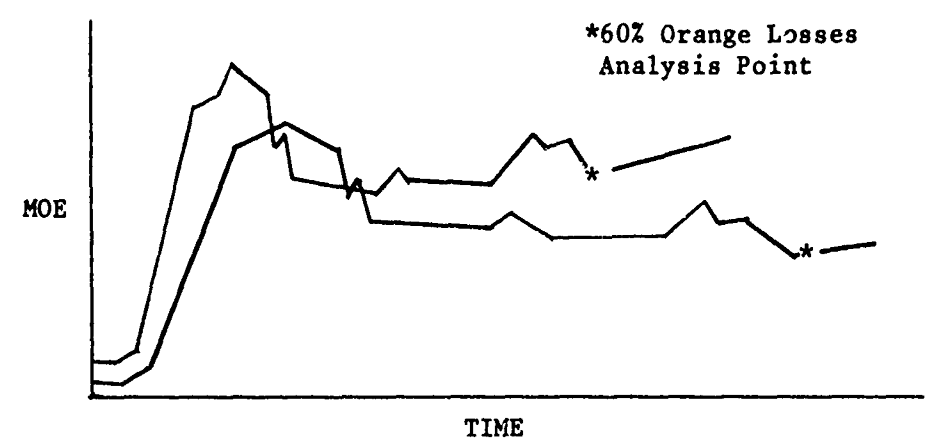


FIGURE 6 MOE PLOT OF DIFFERENT MOBILITY AND EFFECTIVENESS ALTERNATIVES

6.3.1. Equal Time Analysis Point

An examination of Figures 5 and 6 shows that an equal-time analysis point may be appropriate for the case with equal mobility alternatives (Figure 5) but not for the case where the alternatives have different mobilities (Figure 6). Some of the other considerations for the selection of an equal-time analysis point are that all major weapons have been exercised, the defenders have fallen back and are fighting from their final defensive positions, or the attacker is within a specified distance from the defender. The point should not be selected at or subsequent to the time of collapse of either the offensive or defensive forces. The collapse situation is discussed in subparagraph 6.3.3.

6.3.2. Equal Percentage Losses Analysis Point

An appropriate analysis point for the case with different mobility alternatives might be at a point where an equal percentage of either ORANGE or BLUE losses occur. The same general principles pertaining to the selection of an equal-time analysis point, discussed above, would also apply to the equal-percentage-losses analysis point. An example of this type analysis point is annotated in Figure 6. Values that have been used in many recent studies vary between 60 and 80 percent losses.

6.3.3. End of Battle Analysis Points

Two force exchange ratio time plots are shown in Figure 7. One trace is for a force equipped with an effective alternative and the second is for the same force equipped with a relatively ineffective alternative. These situations are annotated in the two traces. The offense collapses with the first alternative (effective candidate) in the force (Case 1). This is shown on the trace by the rapidly increasing force exchange ratio toward the end of the battle. The defense collapses with the second alternative (relatively ineffective candidate) in the force, (Case 2). This is seen on the plot in the rapidly decreasing exchange ratio. If the analysis points were selected in the collapse region, a manager trying to measure the difference in performance of the forces with these two alternatives would obtain different results. In this situation, another alternate approach would be to determine the percent of losses inflicted on the offense by the defensive force when equipped with its weakest alternative at a point just before the defense

collapsed. This point could then be used as the equal-percentage-offensive losses analysis point for all alternatives. This point is shown on the force exchange ratio plots in Figure 7. A major factor that must be taken into consideration in this situation, when considering the results, is that one alternative present in the force provided a win while the other alternative present in the force resulted in a loss.

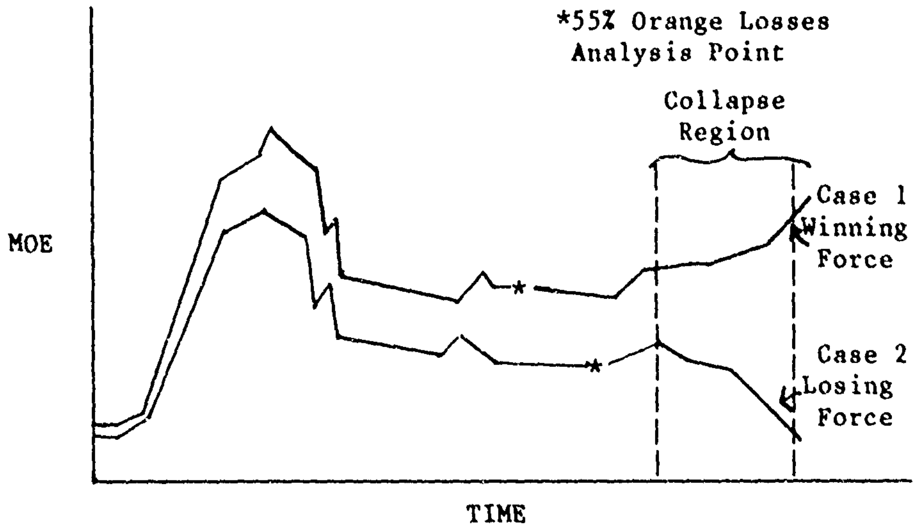


FIGURE 7 MOE PLOT OF WINNING AND AND LOSING FORCES

6.4. Echelon Defense Analysis Point

Another situation that sometimes arises is where the defense is organized in echelons and a "fight and die in place" philosophy is adopted. In this instance, multiple analysis points are usually used, i.e., one for each echelon. Some analysis points that have been used in the past for this situation are the times at which each of the initial echelons were overrun. The final analysis point was a percentage-losses-type analysis point for the force occupying the final defensive position.

6.5. Review of Analysis Points Used In Recent Armor Studies

A summary of the analysis points that have been used by study managers in recent armor studies is shown in Table 9.

TABLE 9

SUMMARY OF ANALYSIS POINTS

FROM RECENT ARMOR STUDIES

	EQUAL TIME	60% TO 80% ORANGE BLUE LOSSES	MISSION COMPLETION
TANK COEA		X	
APC COEA		X	X
AT VEH. COEA	X	X	
APC STUDY 1	X	X	
APC STUDY 2		X	
TANK COEA UPDATE	X		X

An examination of the table shows that a variety of analysis points have been used. There appears to be no trend or hard and fast rules; instead, each point is chosen depending on the specific situation and the considerations discussed.

7. SELECTING COST ANALYSIS METHODOLOGY

7.1. General

The COEAs and most studies include a cost analysis unless they are extremely narrow in scope. This analysis involves computing and comparing the costs for the alternatives to determine possible differences. The costs are important since new systems or improvements to existing systems have cost impacts and must compete with other projects for defense funding. This is also true for alternative force structuring. The types of cost analyses that are available to study managers and used in these studies are life cycle costs (LCCs) and force costs. Both types are discussed below.

7.2. Life Cycle Costs

The LCCs of alternative systems include all expenditures necessary for acquiring, fielding, and sustaining the operation of the system for its assumed life. This includes the cost for research and development, acquisition, and operating and support costs. The LCCs are computed for a fixed period of time, but excursions are normally performed to check the sensitivity of the costs to variations in the time period. The LCCs are useful because they provide the study managers and decision makers with relative comparisons of the cost for buying, fielding, and operating the entire procurement fleet of an alternative for a given time period.

7.3. Force Costs

Force Costs consist of the investment and operating costs associated with achieving and maintaining a combat capability over a specified period of time. The capability is for a combat unit defined by a Table of Organization and Equipment (TO&E). The unit is deployed in a specific operational environment and contains specific items of materiel and operating and maintenance personnel. Force costs are interesting since they provide the decision makers with the cost for total resource implications associated with acquiring and maintaining a units combat capability. An interesting feature of the force costs is that the larger the size of the force that is costed, the less difference there is between alternatives. This occurs because the cost difference between the alternatives tends to be overcome by the additional costs associated with the large organization.

7.4. Review of Cost Methodology Used in Armor Studies

The cost methodology used in recent armor COEAs and studies is shown in Table 10.

TABLE 10

COST METHODOLOGY OF RECENT STUDIES

COEA/STUDY	LIFE CYCLE COSTING				FORCE COSTING
	5 YEAR	10 YEAR	15 YEAR	20 YEAR	
TANK COEA		X*	X*	X	X
APC COEA		X*	X*	X	X
AT VEH. COEA				X	X
APC STUDY 1				X	X
APC STUDY 2	X*			X	X
TANK COEA UPDATE		X*		X	X

*Excursion.

An examination of the table shows that most studies have used both life cycle and force costing methodologies. Also, the life cycle analyses have excursions for the system life time period. The present trend is to develop LCCs based on an assumed operating life of 20 years. Sensitivity analyses of LCCs are also performed to examine the changes in LCCs due to variations in other input parameters such as reliability, availability, and buy quantities. A compilation of the sensitivity analyses that were done as a part of the LCC analyses in each of the studies is shown in Table 11.

TABLE 11

LIFE CYCLE COSTS (SENSITIVITY ANALYSES)

STUDY/ COEA	SENSITIVITY ANALYSIS					
	LIFE PERIOD	BUY QTY	AMMUNITION TRAINING	AMMUNITION WAR RESERVE	EQUAL COST	PRODUCTION RATE CHANGES
TANK COEA	X	X			X	
APC COEA	X	X	X	X		
AT VEH. COEA		X				
APC STUDY 1		X				
APC STUDY 2	X	X				
TANK COEA UPDATE	X	X			X	X

Although force costs are still being computed, they are not viewed with the same importance since the current trend is to no longer combine relative force costs with relative effectiveness to obtain relative worth. (This is discussed below). Although both costs are still computed, the current trend then is for study managers to rely more heavily on LCCs than on force costs.

8. DETERMINATION OF COST/EFFECTIVENESS RELATIONSHIPS

8.1. General

Cost and effectiveness can be computed and the results can be considered separately or combined using cost effectiveness relationships. Two of the methods that have been used in past studies to integrate cost and effectiveness are relative worth and attrition costing. Another related concept is an equal-cost variable effectiveness analysis.

8.2. Relative Worth

Relative worth is the ratio of relative effectiveness (the effectiveness of the alternative normalized to the base case) to relative cost (the cost of each alternative normalized to the base case). Presented in equation form, it appears as

$$\text{Relative Worth (RW)} = \frac{\text{Relative Effectiveness}}{\text{Relative Cost}} \qquad (7)$$

The costs used in the above equation are the force costs (reference paragraph 7) of the force whose effectiveness was determined in the force-on-force combat modeling. Therefore, a RW figure of merit is calculated for each alternative-equipped force for each scenario in which effectiveness and force costs analyses were performed. An example of RW is shown in Table 12.

TABLE 12

RELATIVE WORTH FOR A LONG RANGE BN TF DEFENSE SCENARIO

ALTER. EQUIPPED FORCE	FORCE EFFECT (LER)	RELATIVE FORCE EFFECT.	20 YR FORCE COST	RELATIVE FORCE COST	RELATIVE WORTH (REF EQUA. 7)
Base Case	3.86	1.0	$15.0M	1.0	1.0
1	4.63	1.2	$15.6M	1.04	1.15
2	5.40	1.4	$16.4M	1.06	1.32

This philosophical objection to RW is that it destroys information, i.e., the effectiveness and cost values. Advocates purport that it simplifies the decision problem, but does so only if a manager is willing to accept arbitrary choices of MOE, analyses points and scenarios.

8.3. Attrition Costing

Attrition costing, when used, is also conducted for each scenario in the study. It uses both the force costs that were computed as a part of the cost analysis and the casualty results from the effectiveness analyses. The costs include the force cost for a specific time period (usually 20 years) and the costs for the combat losses and ammunition consumption that are both determined from the combat model. An example of an attrition cost curve is presented in Figure 9.

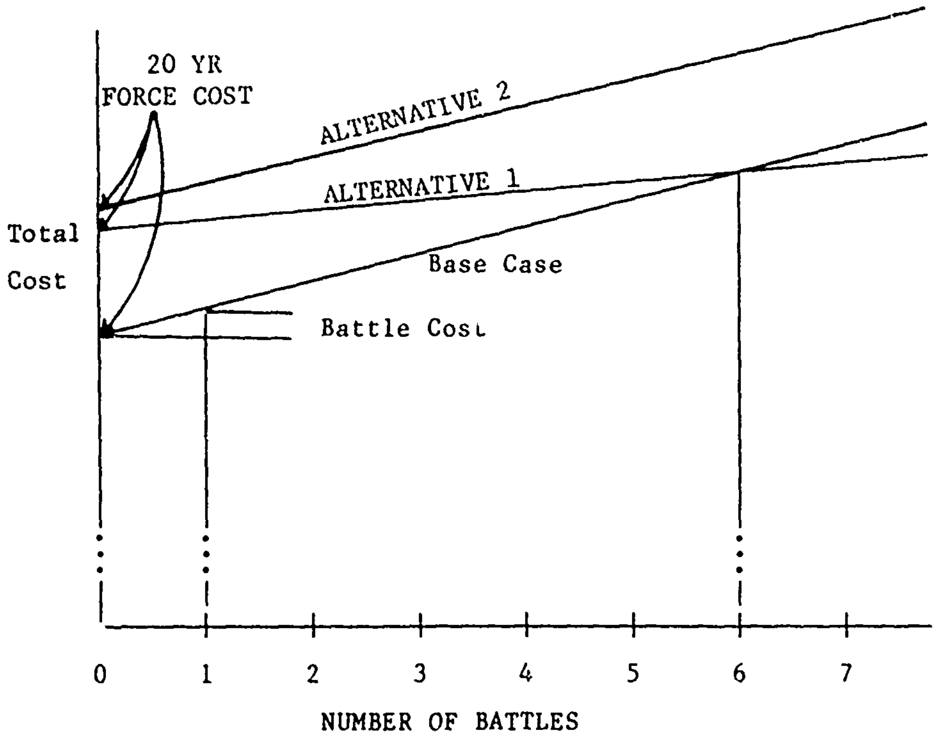


FIGURE 9 ATTRITION COSTS FOR LONG RANGE BN TF DEFENSE SCENARIO

The force costs are used in the attrition cost analysis to identify the absolute cost differences for equipping and sustaining the alternative-equipped forces for the time period. These are the costs which appear for zero battles in Figure 9. Costs to the alternative equipped forces as a result of combat losses and ammunition consumption to attrite a specific number of enemy casualties are then computed for each battle. The cost per battle for each alternative is the

slope of the alternative's attrition cost line in Figure 9. For any number of battles then, the sum of the battle costs for that number of battles and the force costs may be determined for each of the alternative-equipped forces and compared to provide an economic ranking. As can be seen in Figure 9, the force equipped with Alternative 1 is less expensive after the 6th battle than is the base case which was originally less expensive due to its lower original force cost.

8.4. Equal-Cost Variable Effectiveness Analysis

Another cost and effectiveness relationship is the equal-cost variable effectiveness analyses. This analysis is performed by using the costs associated with the most expensive system to purchase additional quantities of less expensive alternative systems. These equal-cost forces are then simulated in identical combat scenarios to determine their effectiveness. Equal-cost forces are then compared in terms of relative effectiveness. These analyses are normally conducted at higher organizational levels where the additional equipment can be incorporated into TO&E units. In the TANK COEA, the analysis was conducted at the theater level where larger numbers of the less expensive tank alternatives were used to form additional battalions.

8.5. Review of Cost Effectiveness Relationships Used In Recent Armor Studies

Results of a survey of cost and effectiveness relationships used in recent COEAs and studies are presented in Table 13.

TABLE 13

COST AND EFFECTIVENESS RELATIONSHIPS

COEA/ STUDY	SEPARATE CONSIDERATION OF COST AND EFFECTIVENESS	RELATIVE WORTH	ATTRITION COSTING	EQUAL COST VARIABLE EFFECTIVENESS
TANK COEA		X	Not Used	X
APC COEA		X		
AT VEH. COEA		X		
APC STUDY 1		X		
APC STUDY 2	X			
TANK COEA UPDATE	X			

As can be seen form the table, attrition costing has not been used in any of the recent armor systems COEAs and studies. An equal-cost variable effectiveness analysis was only used in the basic TANK COEA and has not been used since. The RW was used in most of the studies until the APC STUDY 2 and the TANK COEA Update. The current trend among study managers is to consider cost and effectiveness individually and not rely on a comparison of ratios like RW or other cost effective relationships.

9. SIDE ANALYSES

9.1. General

Separate analyses, apart from the combat modeling effort, are often conducted in armor studies to answer specific questions or provide insights for issues. These side analyses

usually relate to the performance of the alternative systems or impacts on the systems organization.

9.2. Performance Analyses

In the performance-type analyses, the parameters are analyzed and compared separately between alternative systems, whereas in the combat models, all the parameters of the alternative and all parameters of friendly and opposing systems synergistically interact as parts of combined arms teams in tactical scenarios. Examples of subjects for performance-type side analyses are detection, lethality, vulnerability, and command and control features. The separate analyses are useful in that in addition to answering questions and providing insights, they help explain cause and effect relationships imbedded in combat modeling results.

9.3. Organizational Impacts

Analyses of alternative characteristics that have organizational impacts are important since these features are not examined in the combat modeling. Transportability; logistical considerations; reliability, availability, and maintainability (RAM); and training are examples of subjects in this category.

9.4. Review of Side Analyses from Recent Armor Studies

A compilation of the side analyses that were performed in recent armor studies is shown in Table 14.

TABLE 14

SIDE ANALYSES

SEPARATE ANALYSIS	COEA/STUDY					
	TANK COEA	APC COEA	AT VEH. COEA	APC STUDY 1	APC STUDY 2	TANK COEA UPDATE
SYSTEM PERFORMANCE						
1. Fire Control	X	X		X		X
2. TARGET Acquisition	X					
3. Vulnerability	X	X		X		
4. Mobility	X	X		X		
5. Survivability		X	X			X
6. Protection				X		X
7. Capacity				X		
8. Design					X	
9. Ammunition Rate					X	
ORGANIZATIONAL INPUT						
1. Logistics	X	X	X	X	X	X
2. RAM	X	X	X		X	X
3. Transportability		X		X	X	
4. Air Transportability		X	X		X	
5. Training		X	X	X	X	X
6. Risk				X		
7. Technical Risk					X	

This list is provided to show what analyses were conducted as part of the studies. It may be used as a list of topics for consideration. The selection of side analysis subjects must be made by each study manager depending on what is necessary for study completeness.

10. SUMMARY OF STUDY RESULTS

10.1. General

Toward the end of a study, after the effectiveness and cost analyses have been completed, the study manager is confronted with the problem of assimilating and considering the results of all the analyses that were conducted. This step in the study process is particularly important since the summary and overall presentation of results are what decision makers rely on for their deliberations. Various methods which can be used to summarize the results are presented below.

10.2. Scenario Weighting

Weighting is one technique that has been used in past studies to mathematically combine the results of the combat modeling scenarios. The weighting of scenarios consists of assigning weights to the various scenarios which correspond to their expected frequency of occurrence. Weights would then be assigned for offense and defense at both short and long ranges and for different terrain types. The effectiveness results are then combined into a single overall effectiveness number using the weights for each. An example of scenario weighting is presented in Table 16.

TABLE 16

SCENARIO WEIGHTING EXAMPLE

SCENARIO	WEIGHTS*	EFFECTIVENESS RESULTS
Short Range Defense	25%	3.5
Long Range Defense	35%	3.0
Short Range Attack	25%	2.5
Long Range Attack	15%	2.0

Overall Effectiveness = 3.5 (.25) + 3.0 (.35) + 2.5 (.25) + 2.0 (.15)

2.86

*Hypothetical Weights

The objection to weighting is that advocates might use this technique to produce a generalized measure of relative effectiveness and erroneously use the abstract value in force optimization techniques.

10.3 Rank Order Charts

The use of ranking charts is another technique used in recent studies to summarize the results of effectiveness, cost, and side analyses. A sample ranking chart for combat modeling effectiveness results is presented in Table 17. Separate tables are used for the cost and side analyses results.

TABLE 17

EFFECTIVENESS RANKING CHART

ALTERNATIVE IN FORCE	SHORT RANGE DEFENSE	LONG RANGE DEFENSE	SHORT RANGE ATTACK	LONG RANGE ATTACK
A	1	1	1	1
B	2	2	3	1
C	3	3	2	2
D	3	3	3	3

Only significant differences are noted in this type of table and alternatives whose performances are not significantly different are given equal ranking. From Table 17, for example, the performance of Alternatives A and B for the long-range attack was significantly better than that of the other alternatives, but neither's performance was significantly better than the other; consequently, both were given a ranking of one for that scenario. This type of chart is useful in that patterns may be visualized more readily than from a table of numerical results. A table of numerical results from which the rankings were obtained should always immediately precede the ranking chart so that the readers who would like to see the actual numbers on which the rankings are based may do so.

10.4 Review of Summary Techniques Used In Recent Armor Studies

Techniques which have been used in recent armor studies to summarize and present results are presented in Table 18.

TABLE 18

SUMMARY TECHNIQUES USED IN RECENT ARMOR STUDIES

COEA/STUDY	SCENARIO WEIGHTING	RANK ORDER CHARTS	SEPARATE CONSIDERATION ONLY
TANK COEA	X		
APC COEA		X	
AT VEH COEA			X
APC STUDY 1			X
APC STUDY 2			X
TANK COEA UPDATE			X

The general trend among study managers is to present all results for the reader's scrutiny. The trend is against weighting of scenarios as the weights are arbitrary educated guesses or are, at best, derived using Delphi techniques. Techniques like rank order charts are permissible but only when the actual data results are readily accessible.

COST EFFECTIVENESS STUDY OF THE ARMY'S TRAINING EXTENSION COURSE (TEC) PROGRAM (ATEC^D)

LTC THURSTON E. PIKE

Army Training Extension Course Program
Training Programs Directorate
Army Training Support Center
Fort Eustis, Virginia 23604

ABSTRACT. This paper presents the methodology and results of a Cost Effectiveness Analysis that has been completed on TEC. This analysis correlates the development of training materials to cost effectiveness. This paper will describe the evaluation of the cost effectiveness of empirically designed training materials against the cost effectiveness of training soldiers in the conventional mode. Specific areas addressed are:

1. Effectiveness/Retention of Instruction

2. TEC Usage

3. Relationship of TEC Use to Skill Qualification Test Results

4. TEC Costs

5. Use of TEC in TRADOC Schools

A summary of how the above areas were correlated is also presented.

A brief history of the TEC Program is given as background information.

1. INTRODUCTION

The US Army's Training Extension Course (TEC) Program had its origin as a result of a 1971 report by a special Army task force, the Board for Dynamic Training. This board cited several problems with Army training. As a result of this report experimental work was begun to explore the concept of exporting products to train soldiers of both the Active and Reserve Components in their individual skills in the unit environment. Due to the successful outcome of these tests a decision was made to expand the program and TEC was born.

Since its inception, the TEC Program has been subjected to a great deal of research. This paper presents a description of the procedures used in the evaluation of the cost effectiveness of TEC performed by the US Army Research Institute in November 1977.

Before examining this cost effectiveness study, it is necessary to present an overview of the TEC Program as it exists today. The program is managed centrally from the Training Programs Directorate, Army Training Support Center at Fort Eustis, Virginia. Here, under the direction of the Program Manager, policy is formulated and resources are allocated to the Army schools and agencies which develop TEC materials through civilian contractors. Training materials are produced following a systems approach called Instructional Systems Development.

TEC MATERIALS

- CRITERION - REFERENCED
- INDIVIDUALLY PACED
- PERFORMANCE ORIENTED
- INTERACTIVE
- VALIDATED
- MULTI-MEDIA

Fig. 1

These materials called kits or lessons are:

1. Criterion - Referenced. They teach specific job related tasks to a specified level of performance.

2. Individually Paced. They allow the learner to progress at his own rate and repeat instruction when necessary.

3. Performance Oriented. They are aimed at hands-on skills which the learner must perform.

4. Interactive. They require the learner to actively respond to the instruction presented.

5. Validated. They are tested on soldiers in order to guarantee that they teach.

6. Multi-Media. Training materials may be developed in an audiovisual, printed or audio only format, following a specific media selection model.

Each of these kits is self-contained. By the end of FY78:

END OF FY78
- 900 KITS PRODUCED
- 2,250,000 COPIES DISTRIBUTED

Fig. 2

Two and a quarter million copies of 900 kits had been distributed to all units of the Active and Reserve Components of the Army. Seventy-five percent of these kits were in the audiovisual format. They consisted of a filmstrip and a synchronized audio tape with a stimulus time of approximately 25 minutes and a response time of 50-60 minutes.

A great deal of manpower and money is necessary to produce such a volume of high quality training material. The following charts illustrate how the program has grown.

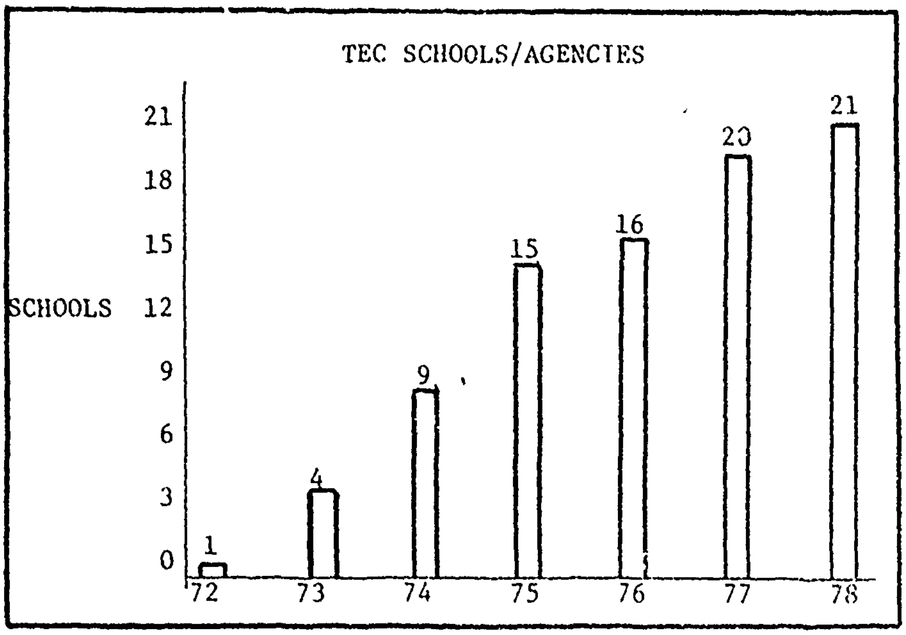


Fig. 3

These schools and agencies provide the subject matter expertise and Government furnished materials necessary for the contractor to produce the lessons. TEC Managers at these schools, together with Project Officers at the Army Training Support Center oversee the development efforts of the civilian contractors.

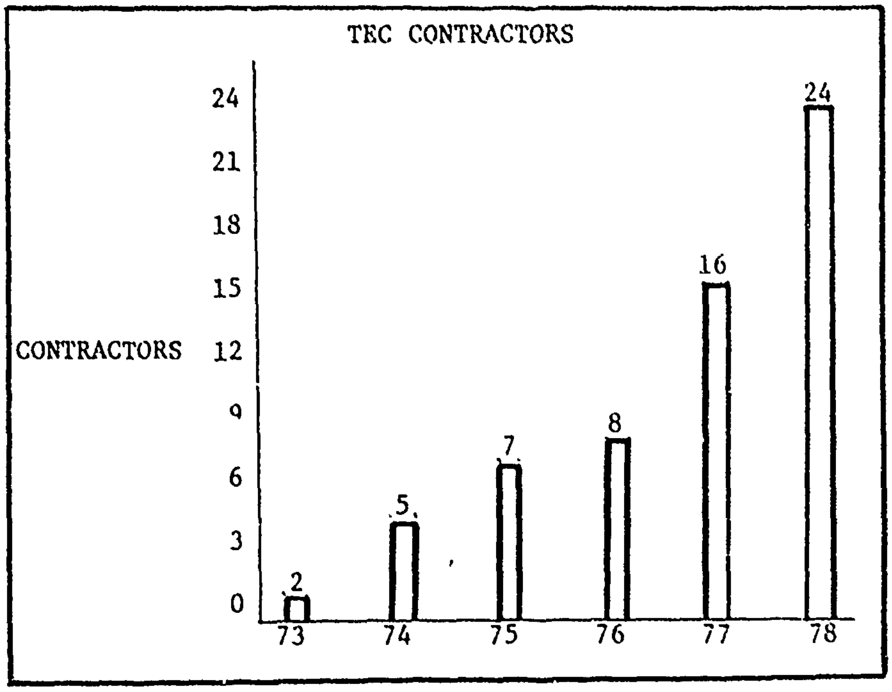


Fig. 4

These contractors, following the specific guidance of the schools, the Program Manager and Government contracting offices, develop and reproduce the majority of all TEC instructional materials.

Funds to support civilian contracts and TEC efforts at the schools are budgeted for by the Training Programs Directorate. The Program Manager exercises complete control and authority over all funds.

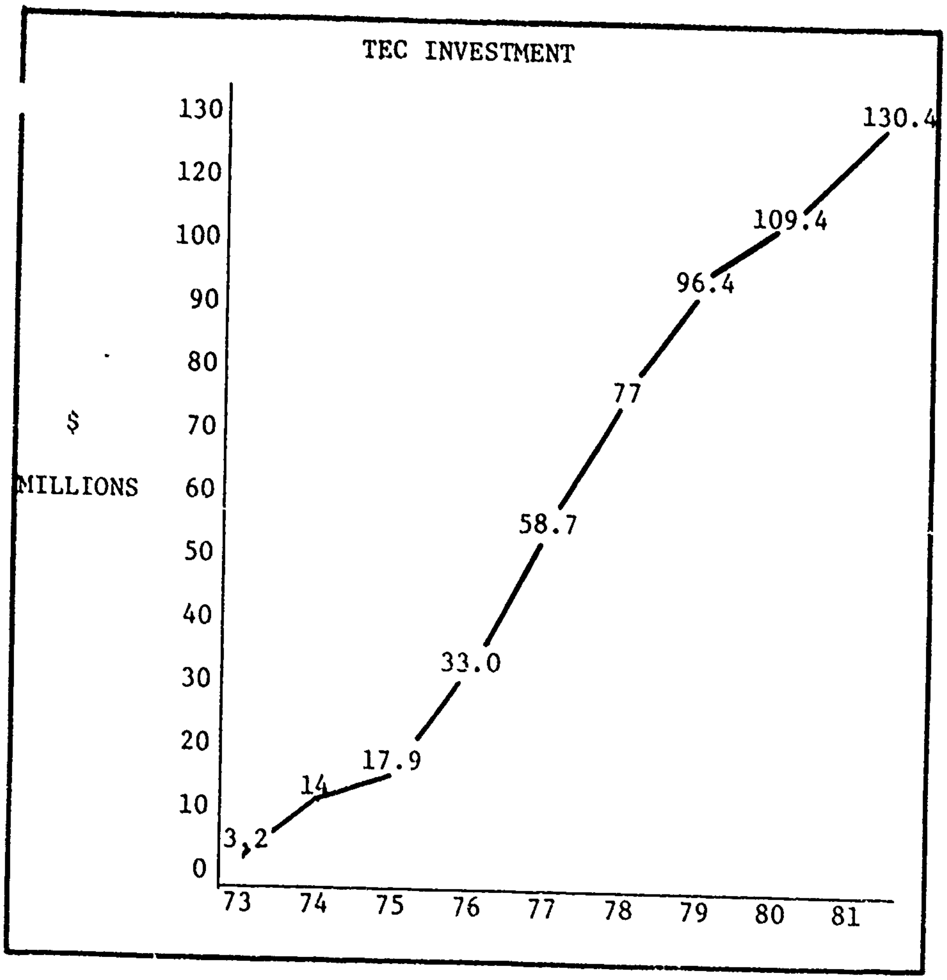


Fig. 5

These investments of men and money have provided the resources to develop 3,429 individual kits.

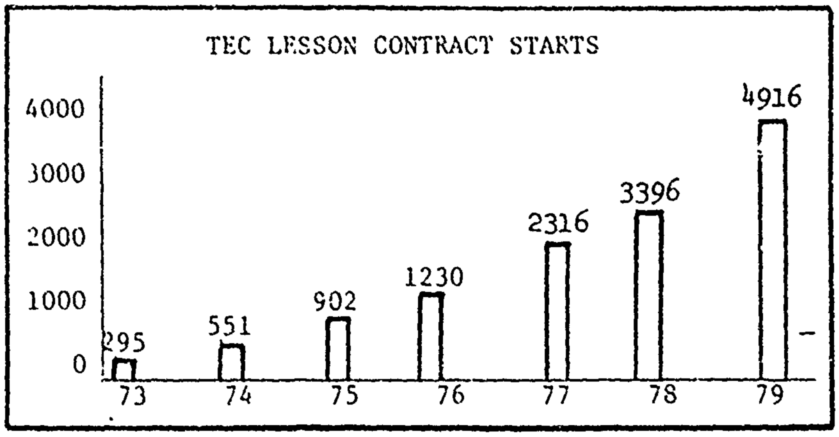


Fig. 6

The TEC Program is now approaching maturity. The development rate is leveling off at 950 kits per year. The required number of copies of each single lesson is determined by the size of the target audience. The average number of copies now required is 2500. This will result in approximately 10 million copies of kits in the hands of our soldiers by the end of 1980.

A program so vast presents many unique management problems. Among them is the problem which this paper addresses - How does one meaningfully evaluate the cost effectiveness of such an undertaking?

2. TEC COST EFFECTIVENESS STUDY - 1977

This concern resulted in the requirement to examine the cost effectiveness of the Army Training Extension Course Program in the area of the individual training of soldiers in critical tasks. The US Army Research Institute (ARI) launched the study. ARI considered several factors that impact on the worth of the TEC Program. Six of these factors were:

1. Effectiveness of TEC lessons in critical task training.

2. Usage levels for TEC in the field and in the service school.

3. Attitudes toward TEC and problems identified in its use.

4. Relationship between TEC and Skill Qualification Test (SQT) performance.

5. Current and programed costs.

6. Example of cost savings and cost avoidance due to TEC.

Several approaches were used in evaluating the above factors. ARI proceeded with the following:

1. A study of the effectiveness and retention of TEC training compared to conventional training.

2. A survey of TEC usage.

3. A study of the relationship between TEC usage and SQT performance for the individual soldier.

4. Identification of the projected costs of the TEC Program over the next ten years.

5. A study of the use of TEC in the Army Service Schools and the resulting cost savings.

A summary of the methodology of each of these five approaches is presented in the following sections.

2.1. Effectiveness Retention Study

The first effectiveness study of the TEC Program was completed in 1975. It reported the following significant advantages for TEC instruction compared to conventional instruction.

FIRST EFFECTIVENESS STUDY

- TEC TRAINED GROUPS WERE CONSISTENTLY SUPERIOR ON PERFORMANCE TESTS
- CONVENTIONALLY TRAINED GROUPS PERFORMANCE WAS MORE VARIABLE.
- TEC WAS EQUALLY EFFECTIVE FOR HIGH AND LOW GENERAL MENTAL ABILITY PERSONNEL.
- TEC TRAINED GROUPS PERFORMED BETTER THAN CONTROL OR CONVENTIONALLY TRAINED GROUPS ON TASKS THAT EMPHASIZED MENTAL ACTIVITIES.
- TEC PROVIDES GREATER INSTRUCTIONAL BENEFITS AT LESSER COST.

Fig. 7

This study, however, did not measure effectiveness using a validated performance test. Nor did it measure subsequent retention of the learning.

The second effectiveness study completed in 1977 corrected both of these weaknesses.

In that study, five TEC lesson series were selected. These series represented lessons specifically designed for each of four major combat arms branches and a series of common lessons. More than 1200 soldiers from both the Active Army and the Reserve Component, representing eight different MOSs, participated in the study. Five experimental groups were created for each of the five lesson series. Two groups received conventional instruction and two groups received TEC instruction. The fifth group was a baseline group.

	CONVENTIONAL		TEC		BASELINE
PRE-TEST	X		X		
INSTRUCTOR TRAINING	X	X			
TEC TRAINING			X	X	
POST-TEST	X		X		
PERFORMANCE TEST	X	X	X	X	X
RETENTION TEST	X	X	X	X	X

Fig. 8

Two of the four groups receiving instruction also received written pre-test and post-test feedback. The performance tests were administered the day following training and again eight weeks later. A simple learning model was applied to the data to correct for prior learning and the effect of the initial test.

A highly summarized version of the results for the Active Army is presented in the following chart.

1977 TEC EFFECTIVENESS-RETENTION
Active Army (Sample Size, 1200)
% Learned & Retained

TYPE OF INSTRUCTION	INITIAL	AFTER EIGHT WEEKS
TEC	38.3	25.5
Live Instructor	27.5	15.4

Fig. 9

The percentage of previously unknown material learned and retained was 38.3% for the TEC trained groups and 27.5% for the conventionally trained groups. Eight weeks later the TEC trained soldiers still performed significantly better. Similar results were obtained in the Reserve Component.

2.2 Usage Survey.

ARI completed a study of TEC usage in 1977. This study included questionnaires and interviews with a large, world-wide sample of users, nonusers, trainers, commanders, and maintenance personnel. Information pertinent to all aspects of TEC usage was gathered from these sources. However, the most useful portion for this cost effectiveness study was the actual quantitative measure of usage that was obtained.

The data were gathered from approximately 100 Active and Reserve Combat Arms Battalions in Oct-Nov 1976. A total of 63,825 uses of TEC was documented during the two month sampling period. The results claimed that an Active Army battalion had an average of 234 TEC impressions per month. A TEC impression is defined as one viewing of a TEC lesson by one soldier. The usage figure for the National Guard was higher - 408 impressions per weekend drill.

2.3. TEC-SQT Study

By gathering data on the use of TEC lessons in sampled units just prior to administration of the SQTs it is possible to determine how the use of TEC relates to subsequent SQT performance. The TEC-SQT correlation phase of the study attempted to determine the degree of correlation between the number of applicable TEC lessons used by the soldier and his Skill Qualification Test (SQT) score. To minimize the effect of extraneous factors in the environment, comparisons were made among soldiers of the same battalions only. Sufficient data for this study was obtained from only two battalions. These battalions showed statistically significant correlation between TEC usage and SQT performance. A current effort is underway to collect this type of data from a larger number of battalions.

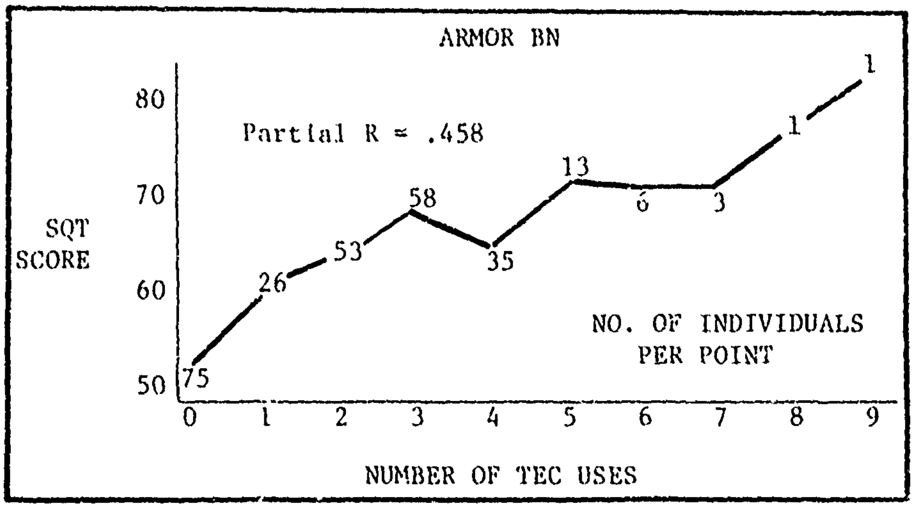


Fig. 10

2.4. TEC Costs Identification

Future costs of TEC were classified as software, hardware, and management costs. Two of these major cost areas were further broken down to basic elements of cost. Software includes the costs of contractual lesson development, contractual reproduction, research and development, in-house reproduction, internal review, quality control, and distribution. Management includes the central operating budget and some personnel costs.

The program consists, economically speaking, of procuring and fielding hardware, protecting the hardware investment through maintenance and replacement, developing new TEC lessons, protecting the software investment through the redevelopment and revision of fielded lessons, reproducing lessons, and distributing lesson copies. Cost assumptions had to be developed based on rather meager data for such issues as hardware maintenance costs, hardware life expectancy, rate of lesson revisions, and rate of lesson obsolescence. The average annual program cost estimate for the 1978-87 period is \$29.9 million. The total cost to provide a lesson copy to a battalion, including a share of the hardware costs, is approximately \$30.00.

2.5. Cost Benefits From TEC Use in TRADOC Schools

Although TEC was not designed nor specifically intended to replace any instructional efforts performed within the resident portions of service schools, it was found that, in many instances, it was so used. This discovery prompted further investigation which revealed other "spin-off" application of TEC material. Since the costs involved appeared to be significant, a method to quantify them was developed.

The impact of TEC use in the schools was categorized into one of the following three areas:

1. Cost Savings. In those instances in which TEC was used to substitute for conventional instruction the costs for instructors replaced and student time savings, where applicable, were computed.

2. Cost Avoidance. In those instances in which TEC development efforts were used as the basis for the development of new in-school instruction the development costs avoided were computed.

3. Remedial/Supplemental Use of TEC. In those instances in which TEC was used to supplement conventional instruction no direct savings nor cost avoidances were computed as this usage did not eliminate nor avoid the expenditure of any resources.

It was found that in 1977 in excess of 2½ million dollars in cost benefits were being derived at the six TRADOC schools surveyed. However, since a large degree of variance existed among the savings at each of the schools, it was determined that adequate projections could not be made from the data. Because of this the factor of cost benefits in the TRADOC Schools was not further integrated in the final cost analysis results.

2.6. Cost and Training Effectiveness Analysis

The individual areas described in the previous sections were associated in two ways to provide evidence of TEC cost effectiveness. The first, the cost benefits quantification approach, is described below. A discussion of the second, a model for individual soldier proficiency, follows.

2.6.1. Cost-Benefits Quantification

The costs and benefits quantification approach attempted to quantify the potential savings in the use of training resources that could accrue from the enhanced training effectiveness associated with TEC use. The focus was on the individual soldier and his tasks. The technique used was the estimation of the impact of the increased proficiency on the performance of specific selected tasks performed during periods of intensive training or combat. Three groups of soldiers were compared. These were a TEC-trained group, a conventionally trained group, and a baseline group that received no special training. The data was obtained based on results from the TEC effectiveness study described earlier. The test results were converted to performance proficiency differences expressed in cost of resources used during an intensive training period.

For example, one of the subject areas studied was the M551 Target Engagement System. The experimental results from the TEC effectiveness study were grouped into the performance measure groups as shown below.

M551 Target Engagement System

PERF MEAS GROUP	TEC INSTR	CONV INSTR	BASELINE	WGTS
Missile reticle	87.3	72.9	63.0	1
Align reticle w/gun/ launcher	56.8	41.6	35.4	2
Align reticle w/coax MG	35.4	16.6	17.3	2
Employing Coax MG	65.0	32.3	36.1	2
Determining Range	34.0	46.5	25.9	2
Weighted Average	52.2	38.5	32.5	

Fig. 11

The weights given to each task reflect the relative importance of that task to operational proficiency. The duty positions within an Armored Cavalry Squadron that required proficiency in the subject area were identified in order to assign a cost of operations for the training period to these duty positions. It was determined that the ammunition costs for the M551 involved in an intensive training period amounted to about $500,000. The next step was to assess the portion of costs that could be saved as a result of differences in training effectiveness between TEC and the other types of instruction. It was assumed that the effectiveness of instruction as measured in the effectiveness study was representative of the effectiveness of instruction for other tasks for the same duty positions. It was also assumed that the operational effectiveness of soldiers at the beginning of an intensive training period was equal to the instructional effectiveness. Finally, it was assumed that the soldier received instruction in the relevant subjects.

It was assumed that learning within the intensive training period occurred at the same rate for all groups of soldiers, and that training continued until actual performance was 80% of perfect performance. The graph below shows the resource savings due to TEC usage under this hypothetical situation. It represents the average results across the given subject areas used in the TEC effectiveness study. The graph shows that a soldier entering an intensive training period immediately following TEC training expected to reach 80% performance with 40.8% less expenditure of resources than the baseline soldier.

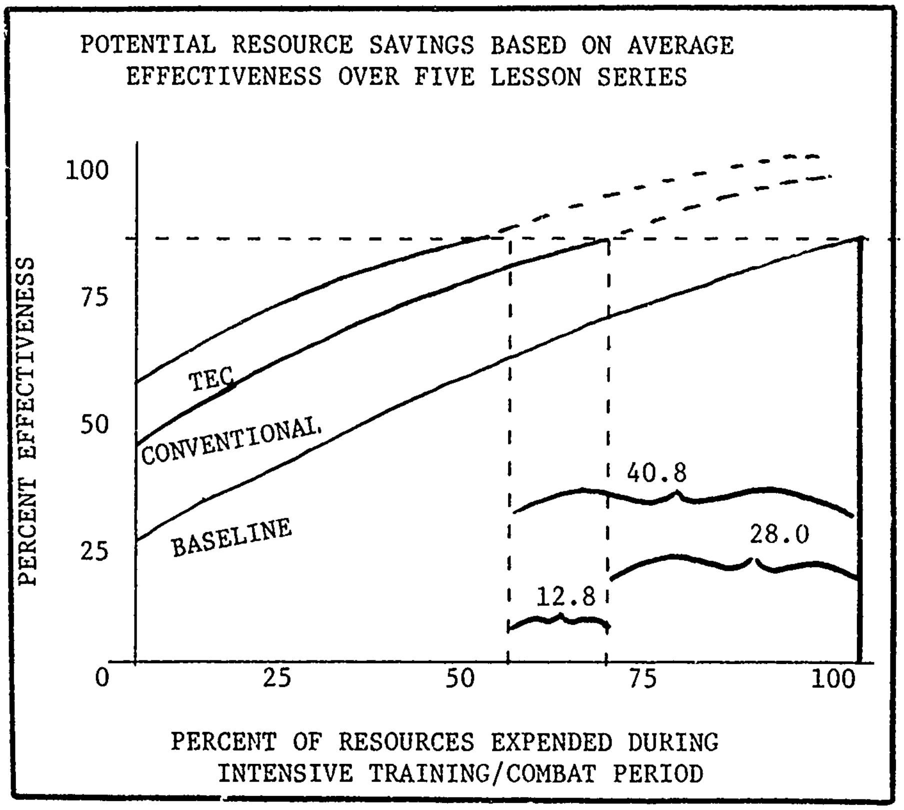


Fig. 12

Mathematically, the percent of resources saved by TEC over baseline is given by the following expression:

$$\% \text{ saved} = \frac{\log (1-T) - \log (1-B)}{\log (1-K) - \log (1-B)} \times 100$$

where T is the initial TEC effectiveness level, B is the initial baseline effectiveness, and K is the terminal baseline effectiveness level.

2.6.2 Model for Individual Soldier Proficiency

The purpose of the model for individual soldier proficiency was to assess the cost effectiveness of TEC by determining the effectiveness gained by the individual soldier from TEC training in his actual unit environment. The basic assumption of the model is that the individual's

average skill proficiency is an additive function of his probable skill at each individual task.

The chart on the following page shows the model and explains its elements.

MODEL FOR INDIVIDUAL SOLDIER PROFICIENCY

$$I_T = \frac{\sum_{J=1}^{S_U} \left(P_{BL_J} + \Delta P_{T_J} \left(R_{F_T}\right)\right) + \sum_{J=S_U+1}^{S} \left(P_{BL_J} + \Delta P_{C_J} \left(PU_{C_J}\right)\left(RF_C\right)\right)}{S}$$

WHERE:

I_T = The individual soldier's average skill across critical tasks at time T.

P_{BL_J} = The soldier's baseline proficiency at critical task J.

ΔP_{T_J} = The gain in proficiency for task J, when TEC is used, compared to P_{BL_J} $(\Delta P_{T_J} = P_{T_J} - P_{BL_J})$

ΔP_{C_J} = The gain in proficiency for task J, when conventional training is given, compared to P_{BL_J} $(\Delta P_{C_J} = P_{C_J} - P_{BL_J})$

PU_{C_J} = Proportion of conventional training actually given for task J per man month.

RF_I = Retention factor of the appropriate value derived from the retention curves generated from the TEC Effectiveness/Retention Study $RF_T \neq RF_C$

S = Number of critical tasks per MOS

S_T = Number of critical tasks per MOS supported by TEC lessons

S_U = Number of TEC supported tasks for which TEC lessons are actually used: $S_U \neq S_T$

Fig. 13

All proficiency elements and criterion factors are taken from the effectiveness research. The estimate of the rate of conventional training was also estimated from training data obtained from the effectiveness sample. The model was applied with current data to derive the estimated average proficiency for an individual infantry soldier at measured rates of TEC usage. This figure was compared to the average proficiency and costs calculated assuming no TEC training. The results showed the TEC dollar holding its own with the conventional training dollar even at low usage levels. This result was based on the comparison of data from both the usage and costs studies discussed above. This comparison showed the cost of a single TEC impression to be $0.39.

2.6.3. Limitations

This study, which managed to correlate the results of four of the five sub-studies, had to contend with many difficulties. One of the most notable was the necessity of dealing with a problem in which both costs and effectiveness were variable. However, the primary problem arose from the need to evaluate the significance of effectiveness differences. This forced the projection of the effectiveness-retention study results to the cost-benefits quantification in the form of the "Intensive Training Period Model" and its attendant assumptions. This, of course, is the fundamental problem in any such study. One can accurately demonstrate program costs, and methods exist to portray gains in knowledge and retention of that knowledge. It is the placing of a dollar value on that knowledge that remains as the missing methodology. This was our attempt at a solution.

REFERENCE:

Strasel, H. C.; J. E. Holmgren; J. Bercos; J. Shafer: and R. Eakins, TRAINING EXTENSION COURSES (TEC): COSTS AND TRAINING EFFECTIVENESS, US Army Research Report Draft, November 1977.

KOREAN ARMOR ANTI-ARMOR ANALYSIS (KOREAN 4A) AND A COMPARISON OF BATTLE ANALYSIS METHODOLOGIES

BERNARD J. DUNN
The BDM Corporation
7915 Jones Branch Drive
McLean, VA 22101, USA

EUN SANG WON
OR/SA Division
DCS, Plans and Comptroller
HQ, Republic of Korea Army

JURI TOOMEPUU
HQ, JUSMAG-K
APO SF 96302

ABSTRACT. The Korean 4A study was undertaken by a joint ROK/US study team to provide quantitative data for anti-armor battle planning and to transfer to the ROK Army the Battle Book Calculus methodology. The paper gives the background for the study effort, detailed description of the methodology, and proposes further applications and improvements. It also compares the problems and advantages associated with manual, computer assisted, computer-player interactive, and purely computer war games.

1. INTRODUCTION

1. 1. Background

The joint ROK/US Korean 4A was conducted during 1 May to 30 September 1978, at the Joint United States Military Assistance Group-Korea. The study was initiated at the request of CINC UNC/USFK/EUSA, and sponsored by the Korean Minister of National Defense, the Chief of Staff of ROKA, and the Director of the US Defense Advanced Research Projects Agency (DARPA). The study was managed by the BDM Corporation, which also provided the study methodology, digitized terrain data of Korea, and the bulk of the analytical effort. Three Korean analysts participated, two from ROKA and one from JCS. JUSMAG-K OR/SA Advisor provided overall coordination and also day-by-day support with the analytical effort.

1. 2. Korean 4A Purpose

The primary purpose was to provide quantitative data for battalion level anti-armor battle planning, to provide inputs for higher level battle simulations, and to transfer the analytical methodology, known as the Battle Book Calculus, to the Republic of Korea. It was also intended to use the method for training of military commanders and staff officers.

1. 3. Scope of the Korean 4A

In order to address the varying types of armor/anti-armor battles that could be fought in Korea, and the current and future threat characteristics that might be encountered in these battles, a spectrum of battle areas, weapon types, and battle conditions were examined. Those weapons that contribute most to the medium and long range armor/anti-armor battle, namely tanks and long range anti-tank missiles and guns, were primarily played. Effects of close air support and artillery we also considered.

1. 4. Study Plan Overview

The Korean 4A Study Plan was structured to accomplish the analysis according to major tasks. Task I encompassed an analysis of potential battle areas and selection of the battalion level battles to be played. In Task II, weapon characteristics were analyzed and kill rates calculated for each important armor/anti-armor weapon pair. The dynamic combat play was conducted for each battle in Task III, including sensitivity analyses of force structure and the effects of combat environments.

2. THE BATTLE BOOK CALCULUS

2.1. Methodology Overview

The advances in deployed weaponry associated with land combat have increased the tempo of modern warfare. Recent combat experience has demonstrated that the high engagement rates and high exchanges of losses characterize the modern battlefield and that speed and firepower are key factors in combat operations. At the same time, the complex digital models which have been used to analyze combat in the past have shortcomings which are serious when used to address combat of this type. Principal among these shortcomings are "invisible interiors" and inflexibility.

The "invisible interiors" deny the analyst the opportunity to observe the way in which the combat variables interact, and the inflexibility of these models prevent the analyst from conveniently making appropriate model changes in the course of analysis.

To overcome the difficulties of the complex digital simulations, a methodology has been developed which minimizes complexity and emphasizes a map-based and template approach which is a more tactically natural technique. This method was developed by The BDM Corporation for use in the US V Corps in Europe and it was then refined for use in the Korean 4A Analysis. This method has become known as the "Battle Book Calculus" methodology because it was used as part of the Commander's Battle Book introduced in V Corps. This chapter presents the methodology in the form in which it was used by military operations analysts to assess armor, anti-armor combat in the Korean 4A Study.

The method is map-oriented, and, while it can be automated, it is deliberately a manual method. By means of a minute-by-minute analysis of the unfolding battle displayed on a map, the analyst obtains direct visibility into the battle dynamics and interactions and has the opportunity, at each minute, to intervene and exercise tactical judgment, for example, to adjust fire distribution or vary the maneuver plan.

2.2. Preparation of Materials

The first step is the selection of a map of the battle area. The scale may be 1:50,000, but scales of 1:25,000 or 1:10,000 are more convenient and provide better resolution.

The second step is to overlay a clear sheet of acetate on the map and locate on it the friendly weapon positions and plot the enemy routes of advance. Figure 1 shows a sample

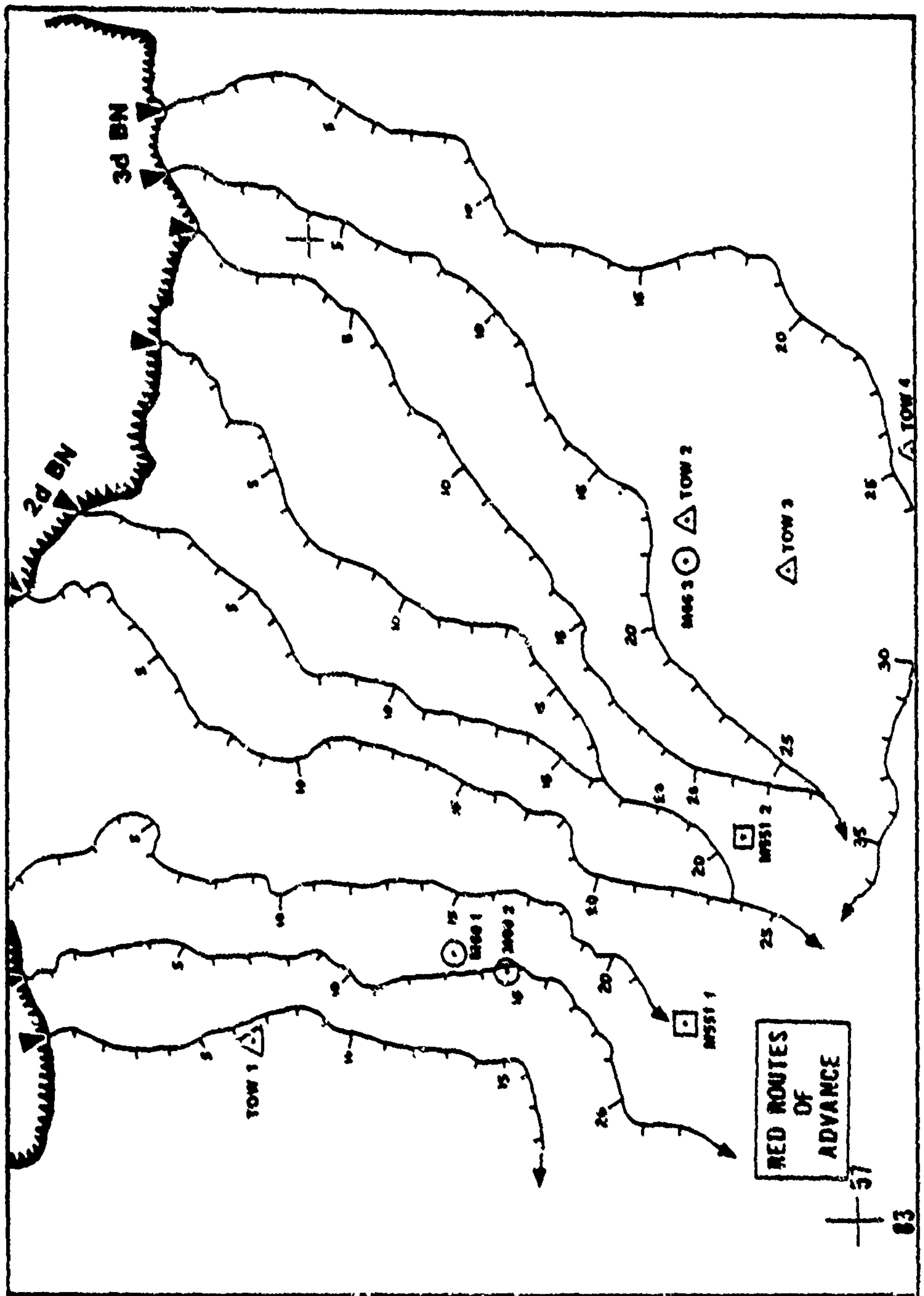


Fig. 1

Map Overlay With Routes for Each Attacking Company and Defensive Weapon Position

case which will be used to illustrate the analysis methodology. The defensive weapon positions and nine attack routes of advance are shown; one for each company (10 tanks) of each of the three battalions of an enemy regiment. One-minute time interval marks are placed on each route to indicate the rate of advance of attacking weapons. In Figure 1 the attack advanced at 12km/hr or 200 meters/min. Variable speeds may be used for different routes or different parts of a route, at the judgment of the analyst.

The acetate should be removed from the map and duplicated to provide a simple black and white copy showing only the attack routes with friendly weapon positions and eliminating the background clutter of the map. This route map is then used to determine individual weapon engagement opportunities for both attacking and defending weapons.

To determine engagement opportunities, the effects of terrain masking must first be accounted for. Figure 2 represents an intervisibility template for a particular weapon position which is used for this purpose. The shaded areas are regions for which intervisibility with the weapon position does not exist. The zone rings shown in the figure are used to play kill probability as a function of range. Grid square indices are used to orient it properly on the route map. Intervisibility templates are needed for each friendly weapon position and may be obtained from:

(1) A computer code utilizing digitized terrain data (Figure 2)
(2) Analysis of the contour lines on a topographic map
(3) Observation from the actual weapon site in the field.

Figure 3 depicts the acetate intervisibility template positioned on the route map. This shows directly those times and the range zones for each route for which intervisibility exists. At the same time, the exposure of the advancing enemy unit may be estimated. Exposure is defined as the fraction of the enemy unit which, due to intervisibility effects, can be seen by the friendly weapon position. Thus, at a time when the enemy route is fully in the clear, the exposure is unity. When the enemy route runs close to a non-intervisible area, the exposure may be 0.5 or some other value less than unity. Exact determinations of exposure require the use of scale templates of the attacking unit formation, but for most analysis an estimate by the analyzers will be more convenient and provide sufficient accuracy. Estimating exposure is important in that it does not permit all live tanks of an enemy unit to fire on a friendly position when in fact only some fraction would be expected to have a line of sight with it. The time of intervisibility, range zone, and exposure data, determined using the technique illustrated in Figure 3 are recorded for each defensive

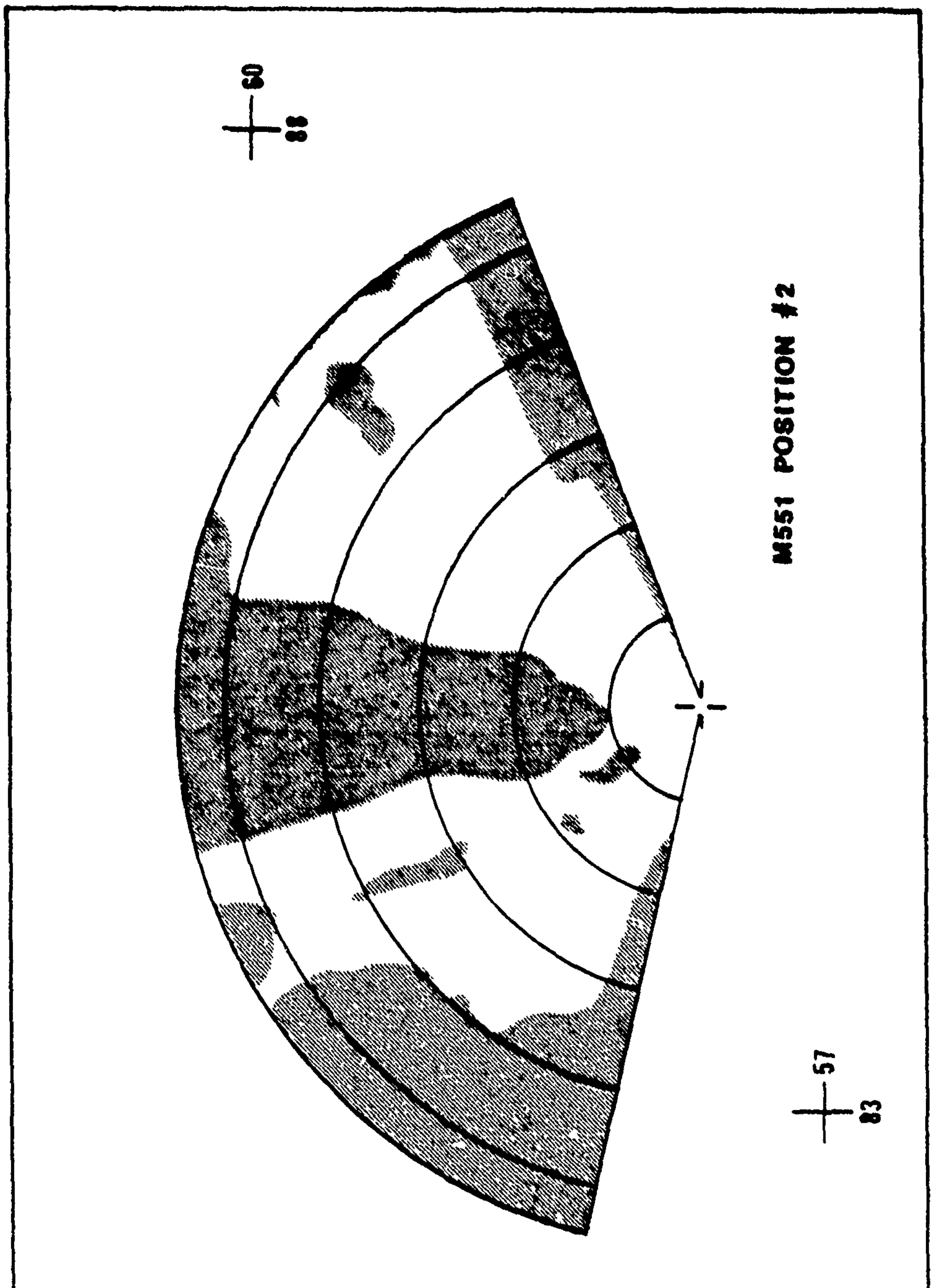


Fig. 2
Intervisibility Template

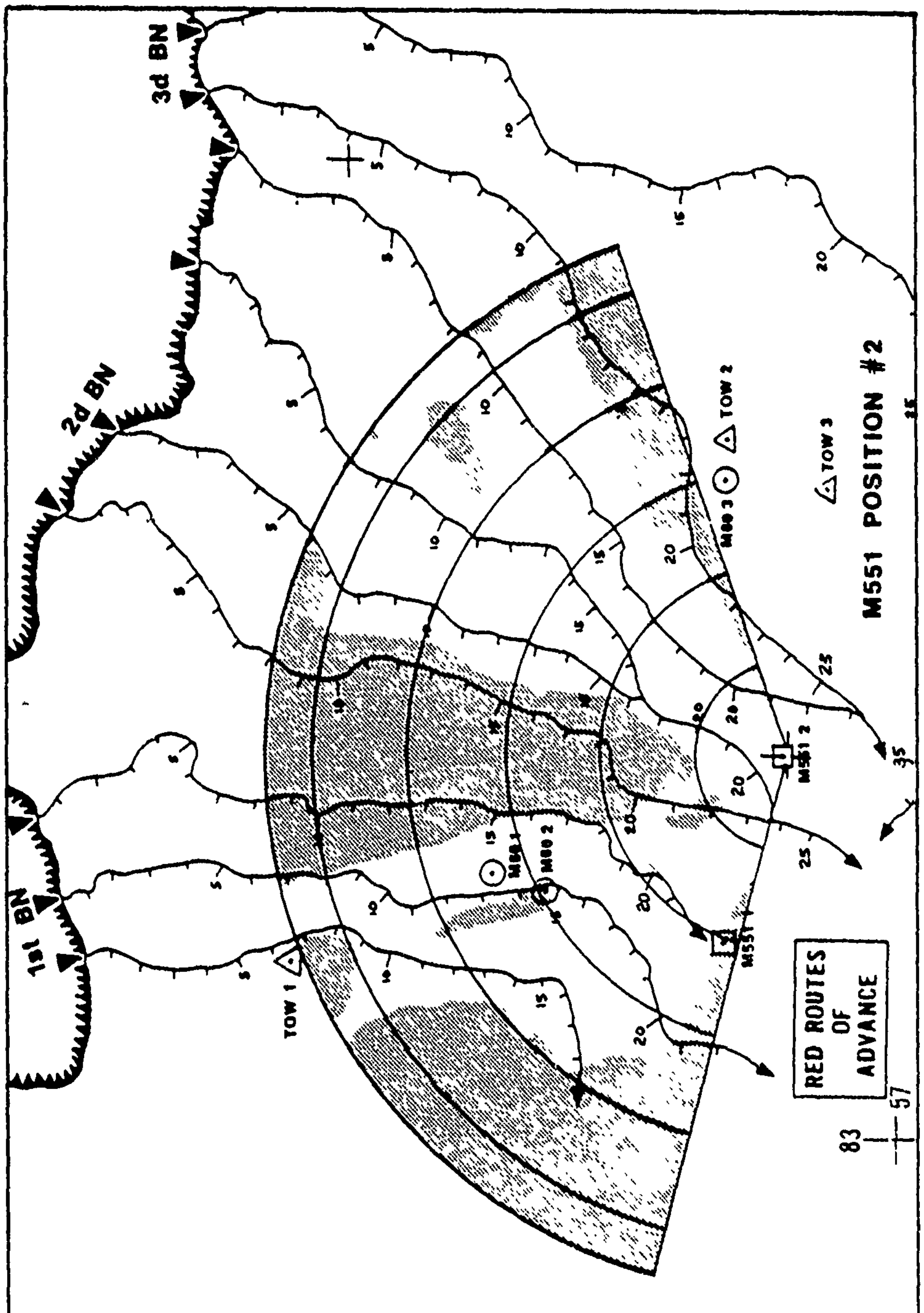


Fig. 3

Intervisibility Template Overlaid on Attack Routes

weapon position. To do this in a form convenient for analysis, a rectified and time-aligned route map is used.

Figure 4 is an example of a rectified and time-aligned route map, referred to as an Engagement Opportunity Diagram (EOD). This is a route map in which all routes are redrawn as straight parallel lines and the time marks all aligned horizontally. Figure 4 shows this for the first battalion of the attacking regiment.

Figure 5 shows the completed EOD with the time, range zone, and exposure data determined by examination of figures of the type of Figure 3 for each defensive weapon position. To the right of each route, a double-headed arrow is drawn to indicate the time intervals during which intervisibility exists for a given friendly weapon position. The double-headed arrow is labeled at its top with the identification of the friendly position. To the right of the double-headed arrow, the estimated exposure is written for each time interval and to the left, the range zone. As an example, Figure 3 shows that for the M551 Position 2 against Route 1, intervisibility starts with the 10th minute, in range Zone 5, at an estimated exposure of 1. Thus, on Figure 5, the double-headed arrow for M551 Position 2 starts at time 10 min, with a 1 to the right and a 5 to the left.

For Route 2, Figure 3 indicates intervisibility beginning at about 9 min, in Zone 5, with an estimated exposure of 0.5, since to the east of the track lies an area of no intervisibility. These observations are reflected on Figure 5, next to Route 2. Note that the coverage of Route 2 for Position 2 is interrupted from time 13 to 15 by a non-intervisible area on the map.

When the construction of the EOD (Figure 5) for all routes and defensive positions is complete, the actual minute-by-minute engagement analysis can begin.

2.3. Engagement Analysis

Figure 6 shows a typical form used to conduct the engagement analysis. This form may be duplicated to provide a copy for each minute of battle. The form shown is for the time interval between 5 and 6 minutes. The engagement assessment of this time interval is conducted in conjunction with the EOD, Figure 5, and the form completed for the preceding time interval. Thus, Figure 6 shows that M60 Position 1 can engage the 1st company of the 1st battalion (Route 1) or the 2nd company of the 1st battalion (Route 2), both in range Zone 3. Tactical judgment must be used to decide. In the example, M60 Position 1 was assigned to the 1st of 1st although the analyst might have chosen the other possibility. This is the same choice the commander in the field must make.

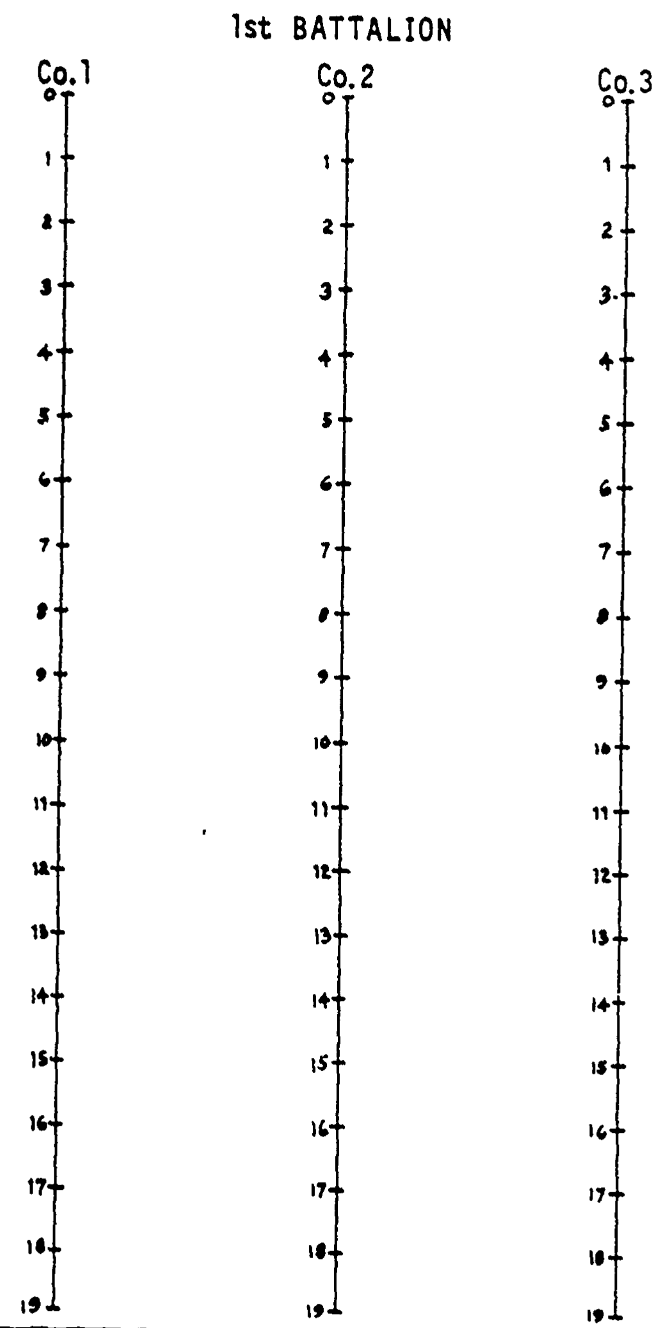


Fig. 4

Rectified and Time-Aligned Route Map

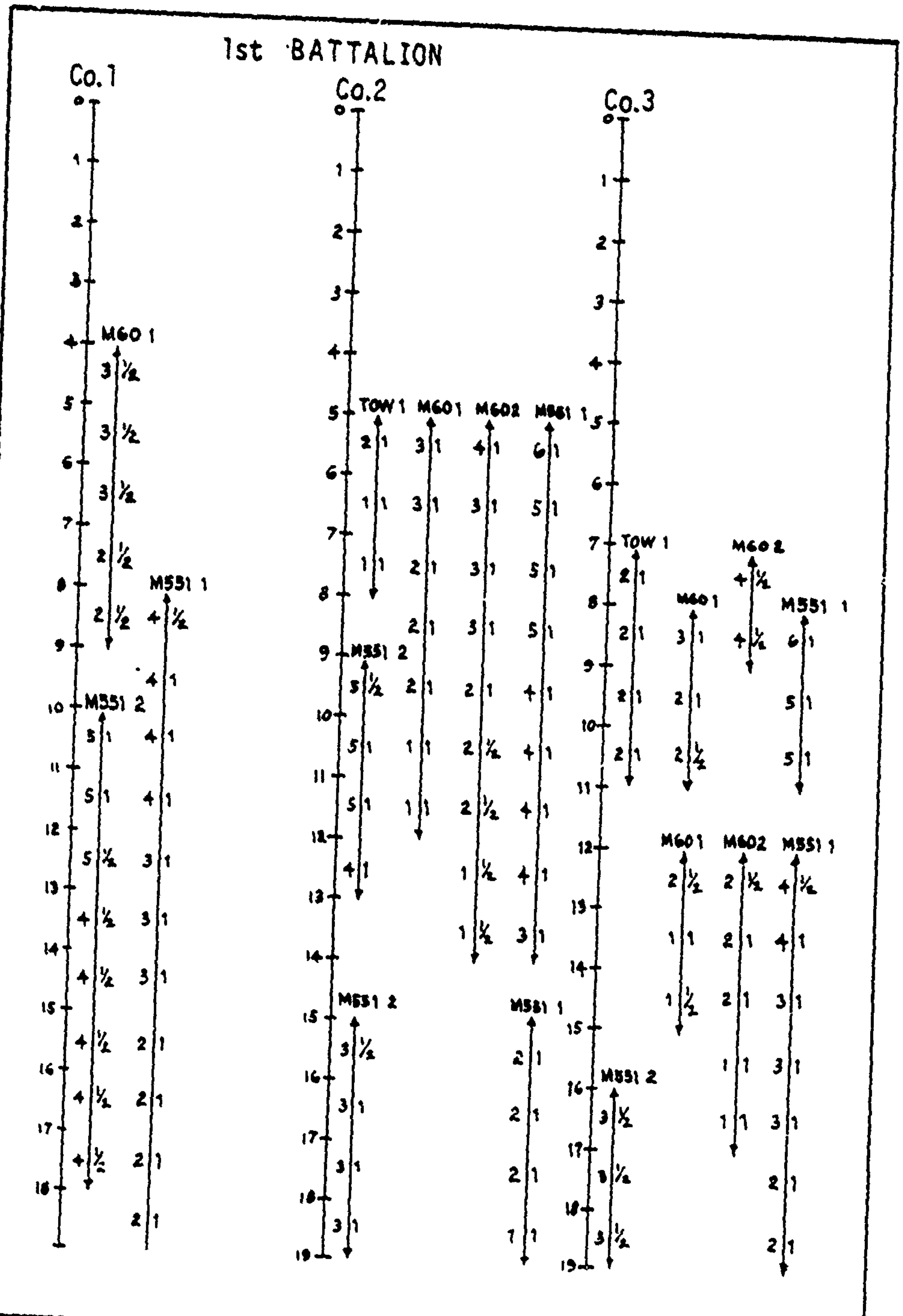


Fig. 5

Engagement Opportunity Diagram With Sample Data

TIME 5 TO 6

ENGAGEMENTS		LOSSES				FORCE ADJUSTMENTS			
BLUE ON RED	ZONE	K.R.	STRENGTH		LOSS	BLUE UNIT	INITIAL STRENGTH	LOSS	FINAL STRENGTH
M60 1 → 1/1		3 ×	4.9		= 1.47	M60 1 :	4.9	− .185	= 4.7
M60 2 → 2/1		3 ×	6		= 1.8	M60 2 :	6	− .2	= 5.8
M60 3 →		×			=	M60 3 :	5	−	= 5
M551 1 → 2/4		15 ×	6		= 0.9	M551 1 :	6	−	= 6
M551 2 →		×			=	M551 2 :	6	−	= 6
TOW 1 → 2/1		.15 ×	2		= 3	TOW 1 :	2	−	= 2
TOW 2 → 2/3		.15 ×	1.4		= .21	TOW 2 :	1.4	− .3	= 1.1
TOW 3 → 1/3		15 ×	2		= .3	TOW 3 :	2	− .176	= 1.8
TOW 4 →		×			=	TOW 4 :	2	−	= 2
RED ON BLUE				EXP	LOSS	RED UNIT			
1/1 → M60 1		.02 ×	8.5	× .5	= .085	1/1 :	8.5	− 1.47	= 7.0
2/1 → M60 2		.02 ×	10	× 1	= .2	2/1 :	10	− 3	= 7.0
3/1 →		×		×	=	3/1 :	10	−	= 10
1/2 →		×		×	=	1/2 .	10	−	= 10
2/2 → M60 1		.02 ×	10	× .5	= .1	2/2	10	−	= 10
3/2 → TOW 2		.02 ×	10	× .5	= 1	3/2 :	10	−	= 10
1/3 → TOW 3		.02 ×	8.8	× 1	= .176	1/3 .	8.8	− 3	= 8.5
2/3 → TOW 2		.02 ×	10	× .5	= .1	2/3 :	10	− .21	= 9.8
3/3 → TOW 2		.02 ×	10	× .5	= .1	3/3 :	10	−	= 10

COMMENTS: TOW 1 withdraws due to approach of 1/1

Fig. 6

Work Sheet With Data, Minutes 5 to 6

The kill rate (KR) for weapons engaging targets is the product of the tactical firing rate and the tactical kill probability (P_k). Both of these are dependent upon weapon system type, target type, weapon and target postures (moving or in defilade), range, and doctrinal and military judgmental factors.

In the present example, which only illustrates the basic methodology, the KRs were, for simplicity, considered range independent, and nothing has been entered in the zone column. In the Korean 4A Study kill rates as functions of range, and weapon and target types, and postures were used.

For an M60 the KR is 0.3 and is entered. The strength of M60 Position 1, as taken from the work sheet for the time interval 4 to 5 minutes, is 4.9 tanks. The product of these numbers, 1.47, is entered in the loss column and is charged against the 1st company of the 1st battalion. This is repeated for each of the friendly units, consulting Figure 5, the EOD. Note that Figure 5 has indicated that M60 Position 1, M551 Position 2, and TOW 4 have no engagement opportunities between time 5 to 6 and, accordingly they can neither engage nor be engaged and attrited in this time interval. Figure 5 shows that TOW 1 has only one engagement opportunity, namely against the 2nd of the 1st, and no judgment is required regarding fire distribution.

For the Red on Blue engagements, a Red unit may engage any Blue position which is firing and with which it has intervisibility. Again, tactical decisions may be required, as indeed they would be in the field. For example, in Figure 6 the 3rd of the 2nd is firing on TOW 2. This is possible since the complete Figure 5, showing all nine routes, shows that intervisibility does exist and TOW 2 (as Figure 6 shows) is firing, although not at the 3rd company of the 2nd battalion.

After the engagements and losses have been determined, the Force Adjustments portion of Figure 6 is completed. Thus for M60 Position 1, the strength at time 5 (from the previous work sheet, time 4 to 5) was 4.9 tanks. This is entered under the "initial Strength" column. Inspection of the Red on Blue engagements and losses shows M60 Position 1 lost .085 to the 1st of 1st and 0.1 to the 2nd of 2nd, for a total loss of .185. This is entered in the loss column. The final strength (strength at time 6) is therefore 4.9 - .185 = 4.7. Note that strengths are rounded to the nearest tenth. The final strengths calculated for one time interval become the initial strengths for the next.

A running display of Red forces, positions, and strengths is obtained by using the actual route map (Figure 7) and updating it after each minute by writing next to the appropriate minute mark the strength of the Red unit at that time.

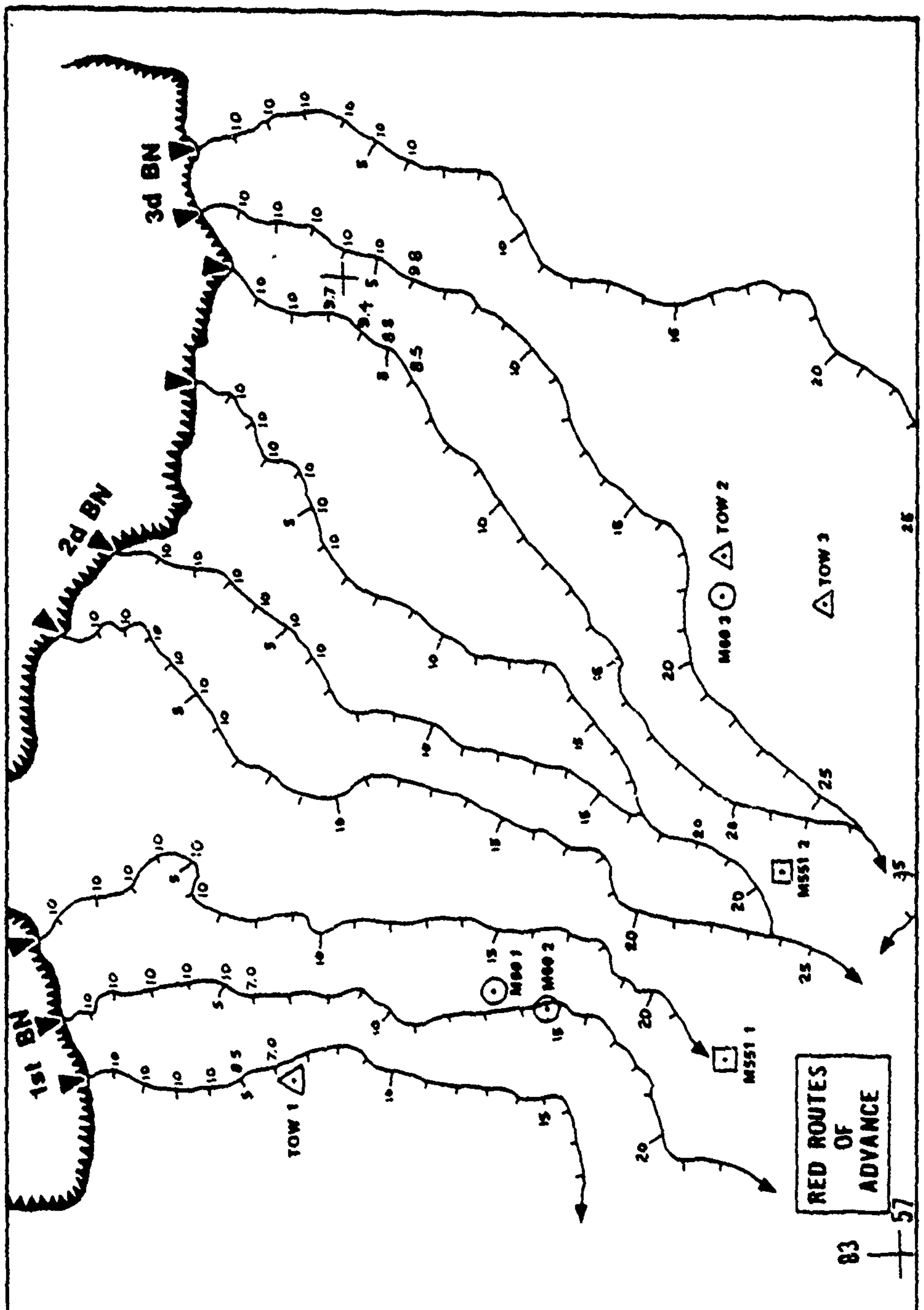


Fig. 7

Status of Enemy Forces, Minutes 5 to 6

Thus Figure 7 shows this update after the calculations for time interval 5 to 6 have been made. On the basis of this situation map, tactical decisions may be made. Thus, the close approach of the 1st of the 1st to TOW 1 produced a decision to withdraw TOW 1. This decision is shown as a comment at the bottom of Figure 6.

Insights with respect to deployment are directly available. TOW 1 was in the battle for one minute before being withdrawn. It was deployed too far forward for a long-range weapon. TOW 4, as it turns out, was deployed too far to the rear, even for a long-range weapon, to get into the battle until quite late.

2.4. Summary of Results

When work sheets for each time interval have been completed, covering the time span of interest, summary charts can be prepared, displaying results concisely and in a form from which insights may be drawn.

Figure 8 displays kills by Blue units. Each Blue unit is listed across the top and time intervals are listed vertically. The kills by each Blue unit for each time interval are entered from the work sheets and when these kills are summed horizontally, the total kills in each time interval are obtained. These can then be graphed as shown to show how the battle intensity varied with time. The summation of the kills per time interval column yields the overall total kills.

The minute-by-minute work sheets may easily be modified to keep a running track of ammunition expenditures. For example, a unit's kills for a particular minute can be divided by the P_k for the appropriate zone to compute the number of rounds expended in that minute. Thus, cumulative rounds fired can be accounted for so that a unit may be removed from battle when it has expended its basic load.

Other summary data may be completed and presented in tabular or graphical forms suited to the purpose of the analysis. The method lends itself readily to adjustments in the minute-by-minute calculations to develop data needed for special analyses. The method can also be adjusted at any time during the battle to alter plans based upon the developing character of the tactical situation. Because the techniques used are largely tactically natural, military personnel using the method can emphasize the tactical play without being distracted by mysterious mathematical or simulation techniques.

KILLS BY BLUE UNITS

TIME	M60 1	M60 2	M60 3	M551 1	M551 2	TOW 1	TOW 2	TOW 3	TOW 4	TOTAL
0-1										
1-2										
2-3							3			0.3
3-4							.285			0.3
4-5	1.5						.255	.3		2.1
5-6	1.47	1.8		.9		.3	.21	.3		5.0
6-7	1.41	1.74	1.5	.9	.9		.165	.27		6.9
7-8	1.38	1.68	1.5	.9	.9			.28		6.6
8-9	1.35	1.53	1.5	.9	.855			.21		6.3
9-10	1.26	1.47	1.47	.9	.81			.21		6.1
10-11	1.2	1.41	1.38	.87	.795			.21		5.9
11-12	1.2	1.32	1.32	.81	.795			.21	.3	6.0
12-13	1.14	1.23	1.23	.81	.795			.21	.3	5.7
13-14	1.08	1.17	1.14	.795	.78			.21	.3	5.5
14-15		1.08	1.14	.795	.78			.18	.3	4.3
15-16		1.02	1.11	.78	.75			.165	.3	4.1
16-17		.99	1.08	.78	.735			.15	.3	4.0
17-18		.96	1.08	.765	.705				.3	3.8
18-19				.765	.69				.285	1.7
19-20				.765	.675				.255	1.7
TOTAL	13.0	17.4	15 5	12.4	11.0	.3	1.2	2.9	2.6	76.3

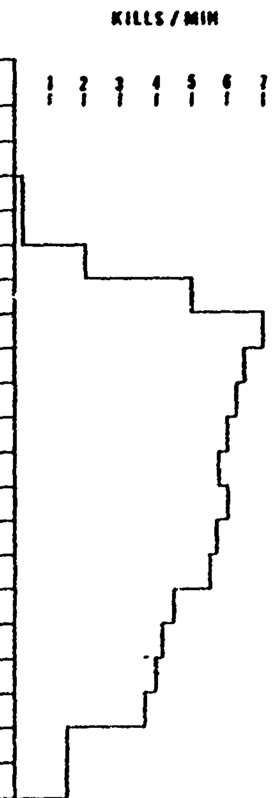


Fig. 8
Kills by Blue Units

2.5. Applications and Proposed Future Improvements

The Battle Book Calculus technique stresses straightforward manual methods. As such, the method can be readily modified and improved. For the Korean 4A Study, it was improved to treat indirect fire artillery attrition, close air support, attack helicopters, and environmental factors such as reduced visibility and smoke obscuration.

It is planned to examine an improvement in the minute-by-minute attrition calculation by using a Monte Carlo-type calculation. In this modification, the weapon time line on the Engagement Opportunity Diagram will be time ticked to show discrete firing times in accordance with the tactical firing rate. At each firing time a random number drawn against the tactical kill probability will decide if a kill is achieved or not. Random numbers may be read from a table or generated by a hand held calculator. This technique may be more realistic in that only integral numbers of fire units on both sides will exist at any time. At the same time, the calculations will be simpler since fractional fire units never appear.

3. COMPARISON OF WAR GAME MODELS

3.1. Trend Toward Complexity

Since the advent of the computer age, the trend in war games used for analyses of the effectiveness of weapons systems and military forces has been toward complexity.

The reason is obvious, warfare is a highly complex interaction of machines, men, and the method of employment of man/machine systems in various physical environments, and computers greatly add to our capability to cope with great amounts of data and a large number of interactions and computations.

As long as the necessity remains for allocation of major portion of our resources for the preparation and conduct of war, there will be talented and well-intentioned critics who can point out factors that are not represented in any given war game, and talented and well intentioned analysts who will attempt to include these factors by desigining war games with greater detail and greater complexity.

3.2. Lack of Validity and Comprehension

It is now becoming clear, that in our most sophisticated computer war games we have reached the point where scientifically valid and up-to-date real world combat or experimenfal results are simply not available to support the

data requirements. This forces us to make an increasingly greater number of assumptions about the natuie of the random processes of war. The complexities of many of the current models preclude the comprehension, by even the most talented analysts, of how the interactions of these influence the results of the game. Also, the validation process is so difficult that most models are superseded by even more complex models before they get validated.

Further, the most important factors of battlefield effectiveness, the intuitive thinking of soldiers and their leaders, their training, experience, morale, and esprit de corps, in short the factors that determine the actions of the man/machine combat systems in continuously changing situations on the battlefield, are not readily quantifiable and quite difficult to model. Because of this, even the most detailed models do not include these human elements of the battle.

3.3. Coping with the Problems of Complexity

Two divergent approaches for coping with these deficiencies in our computer war games suggest themselves.

First, we can continue to attempt to account for an increasingly greater number of factors in the battle by trying to gain a better understanding of these and their interactions, making the effort that is needed to validate the models and the assumptions in them, and by increasing the capabilities of digital computers to the point where they can approximate the processes of thinking.

Theoretical considerations, based on thermodynamics and quantum theory applications to data processing systems, place into question the success of this approach. In a paper intitled "New Optimal Design and Analysis Techniques Using Variational Principles of Nonequilibrium Thermodynamics", Berthold Zarwyn [1] from Harry Diamond Laboratories pointed out that the number of move sequences in a chess game is approximately 10^{120}. Lremermann from the University of California [2] has calculated that no data processing system, artifical or living, can process more the c^2/h bits per second per gram of mass, where c is the velocity of light and h is the Planck's constant. This equates to $2x10^{47}$ bits per second per mass. Based on this theoretical data processing speed, which has not even been approximated by current electronic devices, a straightforward attack on a chess game would require a computer larger than the earth.

3.4. Methods of Decreasing the Complexity.

The other basic approach to coping with the problems of war games is to decrease the complexity of models by replacing the rigid rules and delineations of all possible eventualities required in computerized war gaming with human judgment , by going back to manual, or by using computer assisted or computer-player interactive games.

As pointed out previously, a manual method such as the Battle Book Calculus used in the Korean 4A, can be effectively used for both analytical and training proposes.

Having been closely associated with the Korean 4A effort, I must, however, admit that the most time consuming, tedious, and error prone parts of the study were the numerous hours spent on making and recording simple calculations.

As we know, computers are ideally suited for handling such boring and error prone bookkeeping functions, with great savings of time. Even though it may not be cost-effective to write the necessary programs for a one-time application, the Battle Bock Calculus can be readily converted from a purely manual to a computer assisted method, while retaining the visibility into battle dynamics and allowing for intervention and exercise of tactical judgment by the player.

Distinct advantages over purely computer war games can also be gained by using computer-player interactive games such as IDAHEX. In such games, the more readily quantifiable factors and assumptions known to be reasonably valid are modeled, and provided with the necessary decision rules for computer manipulation, while the tactical decisions are made by the players.

Since human judgment is exercised in the course of conducting manual, computer assisted, and computer-player interactive games, these can also be used for training of military commanders and staff officers, or for a combination of analysis and training.

3.5. Conclusion

In conclusion, it is becoming increasingly clear that the complexities of the latest generation of computer war game models have reached the point where important interactions are being obscured by the analytical process, and further, that adding additional detail and complexity to deal with the random processes of warfare seem to be counterproductive. In any event, the all important human element on the field of battle can not be adquately modeled and played.

The most promising approach for coping with these deficiencies seems to be the use of analytical models that allow human judgment to be exercised at every step in the analytical process.

3.6. REFERENCES.

[1] Zarwyn, Berthold, NEW OPTIMAL DESIGN AND ANALYSIS TECHNIQUES USING VARIATIONAL PRINCIPLES OF NONEQUILIBRIUM THERMODYNAMICS, Proceedings, Fifteenth Annual US Army Operations Research Symposium, Vol II, Fort Lee, Virginia, October 1976.

[2] Bremermann, H. J. OPTIMIZATION THROUGH EVOLUTION AND RECOMBINATION IN SELF-ORGANIZING SYSTEMS, Spartan Books, Washington, D. C. 1962.

COMBAT SAMPLE GENERATOR

(COSAGE)

ODIE B. RICHARDSON, JR., LTC, CM, USA

US Army Concepts Analysis Agency (CAA)
8120 Woodmont Avenue
Bethesda, Maryland, 20014, USA

ABSTRACT. COSAGE is the acronym for a new, division-level, combat simulation named the Combat Sample Generator. COSAGE is a dynamic and stochastic, two-sided, combat discrete event simulation program in SIMSCRIPT II.5. Individual tanks, direct fire, indirect fire, and other crew-served weapons are individually depicted. This model was developed over the past 24 months by CAA and is currently undergoing experimental and test production runs.

COMBAT SAMPLE GENERATOR (COSAGE)

The Combat Sample Generator (COSAGE) Model is being developed by the Concepts Analysis Agency (CAA) to assist in the preparation of more accurate estimates of ammunition expenditures and the loss of personnel and equipment in a combat simulation. In the past, the results of combat on the battlefield were evaluated using several different models, each of which portrayed a different aspect of the battle. COSAGE replaces the following eight separate models:

- Tank/Antitank Model (TATM)
- Infantry Combat Model (ICM)
- HOVER
- HOVARM
- Casualty Assessment Model (CAM)
- Target Acquisition Model (TAM)
- Blue Artillery Model (BAM)
- Red Artillery Model (RAM)

The purpose of the development of COSAGE was to provide a model which integrates the functions previously represented separately. It was also necessary that the model developed be efficient in order to provide a responsive method of analysis--that means the shortest computer run time possible and moderate demands for interaction with gamers.

Following a project definition phase, detailed design of COSAGE began in January 1977 and the model became operational in August 1978. The approach involves a division level simulation with resolution to the platoon level and each crew-served weapon may be individually depicted. The following are the major assumptions on which COSAGE is based:

- It is possible to create stylized combat samples which can be extrapolated to develop meaningful theater-wide campaign-long forecasts.

- Task force organization does not change within a day's combat except in the case of one unit reinforcing another.

- Command and control orders can be prepared in advance to direct the response of maneuver units to foreseeable contingencies.

The assumptions seem reasonable for a 24-hour slice of combat.

COSAGE plays a Blue division against an appropriate Red force. Essential features of the COSAGE methodology are that it is:

- Dynamic

- Stochastic

- Programed in discrete event simulation language without fixed time interval (SIMSCRIPT)

- Dependent on gamers for scenario and orders

The gamers select the forces to be played, deploy them on the map and give them their orders. Orders are of four types: move, attack, defend, and request reinforcements. The gamer can select the level to which the orders apply, e.g., battalion, company, platoon. The move order is described here:

- To specified coordinates or distance/direction

- At specified time or contingency

- Order includes
  - Type of move
  - Mission of unit
  - Next order if
    - No oppostion
    - Opposition and all's well
    - Opposition and in trouble

The attack order is issued with a move order of the absolute type:

- Missions
  - Patrol
  - Probe
  - Attack

- Order includes
  - Next order if enemy disengages

- Next order if this unit disengages

The defend order is similar to the attack order in that it includes the assignment of missions and specifies the next order:

- Missions
  - Delay
  - Defend
  - Ambush

- Order includes
  - Next order if enemy disengages
  - Next order if this unit disengages

The request reinforcements order is used in conjunction with the move order. If a unit makes contact and its strength in critical equipment is below a certain level, it can request reinforcements:

- Request to superior unit

- Must find unit
  - Not engaged
  - Not depleted
  - Without overriding orders

- Order includes
  - Next order when reinforcements arrive
  - Next order if no reinforcements available

The following is a sample order set which could be given to an attacking battalion and a defending company, respectively:

Attacker

1. Move to 16831407 at 3.5 hrs
   probe 60, 2, 3, 4
2. Defend 2, 2
3. Attack 1, 4
4. Reinf 1, 2

Defender

1. Defend 1, 2
2. Move withdraw 5000, 1

The attacker's orders direct him to begin to move at 0330 hours to the coordinates shown with a mission of probe. If he reaches the coordinates without opposition, he is to set

up a defensive position (order 1). If he makes contact and his strength (in critical equipment) is above 60 percent, he is to attack (order 3). If he makes contact and his strength is less than 60 percent, he should request reinforcements. If reinforcements are available, he should resume his mission when they arrive. If the enemy unit is still in proximity, he will attack. If the attack is successful, he resumes his original mission--if not, he defends in place. The defender is directed to defend and, if successful, to maintain that defense. If unsuccessful, he should withdraw 5000 meters and set up another defensive position.

The simulated day of combat usually begins at midnight. Units move at times specified by their orders. Their movements are monitored and when two forces with appropriate orders make contact, a battle begins. A typical battlefield appears as shown in Figure 1.

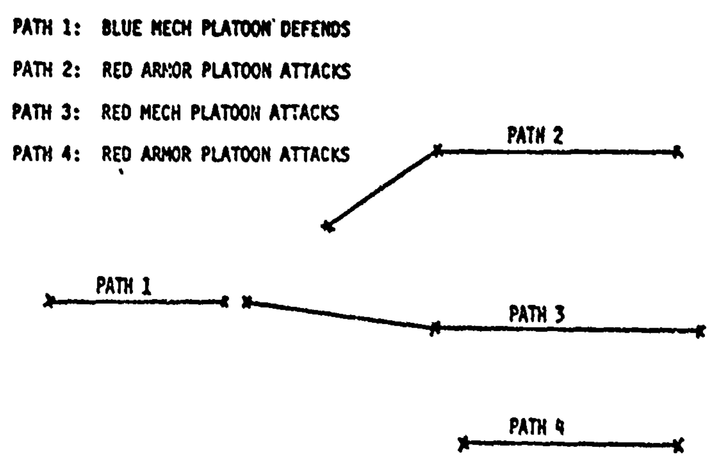


Figure 1. Typical COSAGE Battlefield

This battlefield depicts three Red platoons attacking a Blue platoon. Each platoon has a path and a mission. If the disposition of the forces does not match the predetermined set of typical battlefield situations, then the forces are arranged in a general situation on line, centered, and with headquarters set back a standard distance.

A battle consists of two major processes--movement and firing. Movement takes place along the paths in segments which represent the existence or non-existence of line of sight between the Red-Blue pair of units. The existence and length of the line of sight segment are determined stochastically from data gathered by actual measurements in various locations in Germany and elsewhere. Thus, the battle is played on a statistical representation of real terrain. The existence of line of sight triggers a detection and a detection precedes firing.

The general sequence of events is shown in Figure 2. There is a wait to represent target acquistion time and then the firer selects a target. Time of flight, which can be significant for antitank guided missiles, is explicitly considered. If a kill has been made, determined from single shot kill probabilities, an assessment of damage is made and then there is a wait for both units to recognize the kill. The firer meanwhile reloads and goes back to target selection. The force being fired upon decides whether the proportion of equipment has fallen below the critical threshhold. If so, a withdrawal begins, if not, the battle continues. The face-to-face encounters of small unit engagements are not the only events of interest to the model. The duel of indirect fires is also modeled.

These are the seven basic target acquisition and indirect fire routine functions performed by COSAGE:

- Establish sensor reporting channels
- Establish indirect fire control relationships
- Generate target reports
- Process target reports
- Generate fire missions
- Process fire missions
- Assess effects in status file

The heart of the indirect fire process is presented by the center four functions which model the three links of the indirect fire chain. Sensors are associated with units in the unit status file.

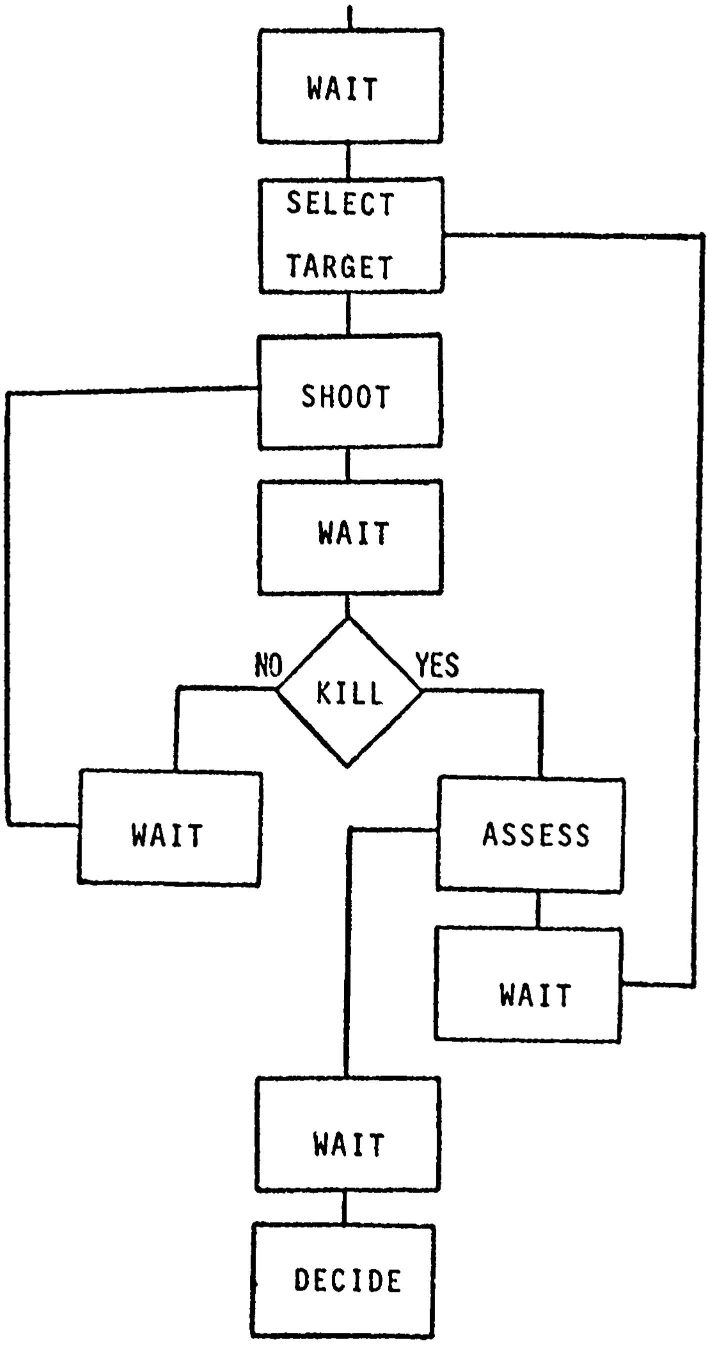


Figure 2. Sequence of Events

Basically, there are two kinds of sensors:

(1) Continuous scan, which monitors the area of coverage for activity continuously when operational and functioning, and

(2) The event activated sensor, which is turned on by sound or by other sensors at specific times.

Sensors have established FDC reporting channels. The artillery force missions and structures are the following: divarty FDC controls all GS and GSR battalions and is the controlling FDC for all DS artillery battalion FDCs. The DS artillery FDCs control their own batteries as well as all batteries of reinforcing artillery battalions, and have the use of GSR batteries unless divarty overrides this control.

When detection is made, a target report is prepared and transmitted through the appropriate fire channels by the sensor which made the detection. Information included in the target report is:

- Estimated location
- Target environment
- Listing of detected elements
- Target movement status
- Reporting unit
- Sensor type
- Start time (preplanned)
- End time (preplanned)
- Target category (preplanned)

At the FDC, the target report is analyzed to determine the estimated category and size; then based on these estimates, the military worth of the target is computed, a duplication check is made against other target reports in the FDC, and the target report is filed in the FDC target report queue. Target reports are removed from this queue for further processing based upon highest military worth upon completion of a target report computation:

- Select next TR from top of TR queue

- Determine desired effects (attack criteria)

- Determine attack method

- Determine batteries available
  - Stationary
  - Operational
  - Not overbooked
  - Not busy at scheduled time (scheduled)

- Rank batteries by proximity to target

- Generate fire missions

- Enter this TR on list of recently computed TRs

- Generate a new target report for higher FDC (if necessary)

After the target report becomes top priority at the FDC, the processing continues. First, the desired effects or "attack criteria" are determined. Then all available batteries are identified; that is, all those batteries under the control of the divarty FDC that are stationary, operational and not overbooked are found. Next, these batteries are ranked based upon proximity to the target.

Now the FDC begins to generate fire missions to be sent to these batteries. For each battery on the list, starting at the top, a fire mission is generated until the attack criteria are met. In this process, first a selection is made between the most effective HE and ICM rounds available at the battery based upon effectiveness. Next, the number of those rounds to fire and the subsequent expected effects are computed. Lastly, the fire mission is transmitted to the battery. Processing of the fire mission now begins at the fire unit. The new fire mission is filed in the battery's fire mission queue where it is further processed based upon priority. If the new mission is of higher priority than the one currently being fired by the battery, that mission is interrupted and the new mission is processed.

Effects are assessed against each specific type of equipment and personnel in a unit based upon the location of the impact point of the fired volleys, and the actual unit location at that time (rather than the location of the unit at the time it was detected). These casualties are then assessed against that unit in the unit status file.

The use of the SIMSCRIPT II.5 simulation language allows for a rapid, dynamic representation of a divison battle. We believe the indirect fire process has been modeled well enough such that we can estimate with confidence the related war reserve requirements for a future conflict where the sides closely adhere to current doctrine and training.

The development of the design of this model is being completed during the current experimental production phase. This spring COSAGE will be used to produce combat samples for sensitivity analysis and comparison with the results of the previous methodology using similar input data.

When the present version of COSAGE has been satisfactorily tested and documented, a study of enhancements will be undertaken. The enhancements below are some of the more important ones to be considered:

- Tactical air
- Air defense
- Obscuration (smoke, weather, dust)
- Minefields
- Electronic warfare
- Ambush

In summary, COSAGE is a high-resolution simulation model of ground combat, programed in SIMSCRIPT, which produces 24-hour stylized samples. Our goal is better data for determining ammunition requirements and losses of personnel and equipment.

AN APPLICATION OF NETWORK SIMULATION TO
FORCE ANALYSIS

Prepared for Presentation
at the
Pacific Conference on Operations Research
Seoul, Korea
23 - 28 April 1979

By

LTC Richard O. Nugent, US Army
LTC Elwyn L. Brown, US Army
LTC Donald J. Berg, US Air Force

of the

Studies, Analysis and Gaming Agency
Joint Chiefs of Staff

December 1978

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

AN APPLICATION OF NETWORK SIMULATION TO FORCE ANALYSIS

LTC Richard O. Nugent, US Army
LTC Elwyn L. Brown, US Army
LTC Donald J. Berg, US Air Force

Studies, Analysis and Gaming Agency
Joint Chiefs of Staff
The Pentagon
Washington, D. C. 20301

ABSTRACT. This paper describes the application of network simulation in the analysis of military force alternatives. A brief review of network concepts and techniques is given. A particular network simulation technique is then discussed, followed by an explanation and demonstration of how this technique can be applied to force analysis.

At present, no widely accepted methodology exists to fit together the many aspects of force analysis. War-game models and theater level simulations provide only a portion of the information which the decision maker, in his evaluation of force alternatives, must weigh. Strategic mobility, force readiness, overflight rights, and the availability of foreign bases for staging and support are just a few of the additional considerations which affect the capability of forces to meet contingency situations. Force alternatives must be evaluated in light of both the capability of allied forces and the ability of the US to provide timely reinforcement.

Network simulation provides a methodology to assist force planners by quantifying the considerations which are part of force determination. The results of war-gaming, together with logistics and mobility analysis, can be pieced together to obtain an evaluation of the various considerations which must be made. Risk analysis can also be performed on the capability of forces to meet contingencies in light of uncertainties associated with threat, readiness, movement, and other considerations.

An example is presented to demonstrate the application of network simulation in the analysis of force alternatives.

I. INTRODUCTION

1.1. Purpose

This paper demonstrates the application of network simulation to military force analysis. This application is made possible by the adaptation of a flexible network simulation technique and computer evaluation program.

1.2. Overview

This paper presents a brief review of network concepts and techniques and then describes a particular network simulation technique. The paper then shows how this technique can be applied to the analysis of military force alternatives. An example is given to demonstrate the application.

1.3. Need

Figure 1 depicts some of the many aspects that are part of the assessment of forces. Force assessment, particularly of corps, theater or larger size forces, is now done using many separate and sometimes unrelated analyses. No one model, methodology, or technique now exists to integrate the many factors that must be considered in assessing the overall effectiveness of a force. An important component of this assessment, that of the combat capability of the force, is often done either using static analyses or large-scale computer wargames. These techniques may require numerous assumptions regarding the specific threat, scenario, and other input data. Further, some aspects that affect capability such as logistics and strategic mobility may only be treated superfically in the analysis methodology. Data may be entered as fixed point input when, in fact, there is considerable variability or uncertainty. Results of analysis are often reported without indicating their sensitivity to variations in assumptions and the input data. It is considered desirable to have a methodology where the effect of assumptions and variability of input data can be examined. This paper proposes the use of network simulation for this purpose.

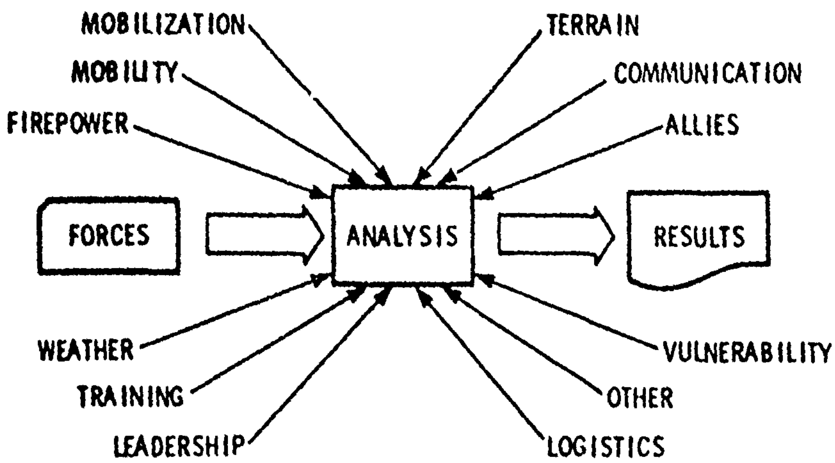


Figure 1

Elements in Force Analysis

2. NETWORK SIMULATION

2.1. Description

Network simulation is a modeling and mathematical technique applied to complex problems to assist the decision maker by breaking down the problem into manageable components. Networks and network analyses play an important role in the Research and Development (R&D) field as the planning and scheduling of project and programs can be readily modeled in network form.

2.2. Elements

The networks discussed in this paper use many of the terms and concepts common to most networks. These include the terms: graph, arcs, nodes, flow and path.

2.2.1. A graph is a drawing of a network using basic symbols for nodes and arcs of the network. The example in figure 2 uses blocks to represent nodes and lines to represent arcs.

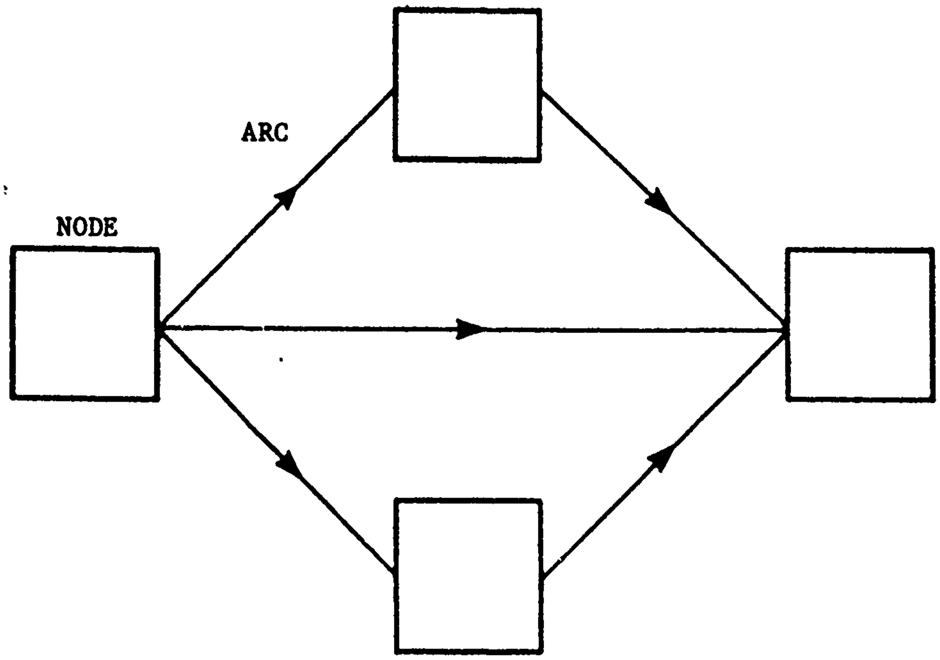


Figure 2

Example of a Graph

2.2.2. Nodes represent events, milestones or decision points, such as the start or completion of an activity.

2.2.3. Arcs represent tasks or activities. Activities consume time and resources and produce some kind of output.

2.2.4. Flow over an arc represents conduct or happening of the activity depicted by that arc. The direction of flow is indicated by arrows on the arcs.

2.2.5. A path is a set of nodes connected by arcs which begin at an initial (source) node and end at a terminal (sink) node.

2.2.6. A network then is a graph in which the arcs are assumed to carry some type of flow indicating the occurrence of a series of activities making up a large project or program.

2.3. Example Networking Techniques

The Critical Path Method (CPM) and the Program Evaluation and Review Technique (PERT) are familiar network techniques used to plan and schedule projects and programs. Recent developments have extended these techniques in order to overcome some of their limitations. These developments have made network simulation a powerful technique involving combinations of functions, probabilities and deterministic values which may defy analytical computations. The network simulation technique explained in the next section is extremely flexible, versatile in its modeling capability and useful in the analysis of complex processes.

3. VENTURE EVALUATION REVIEW TECHNIQUE (VERT)[1]

3.1. History

VERT was developed by Mr. Gerald Moeller of the U. S. Army Armaments Command at Rock Island, Illiniois in 1973. The current version was produced in 1977.

3.2. Description

VERT is a mathematically oriented, simulation networking technique designed to assist in the assessment of risk involved in a project or systems development effort. VERT incorporates stochastic variables, probability of success for each arc and logic on the nodes to reflect decision rules and chance events. In addition to time and cost parameters, performance is included to allow consideration of time and cost tradeoffs for performance and risk. Major features of the technique which make it quite useful in analyses are (1) the transformations which allow modeling parameter values as random variables or functions of variables evaluated elsewhere on network, and (2) the extensive output options and data available.

---

1. The documentation on this technique is contained as part of the computer program and is also explained in an instructional pamphlet entitled: VERT, Venture Evaluation and Review Technique, ALM-63-4080-H1 (B), United States Army Logistics Management Center, Fort Lee, Virginia. The interested reader is referred to either of these because the discussion in this paper is only to provide a feel for the capabilities of the technique.

3.3. Operands

The principle symbolic operators used in VERT are the nodes and arcs.

3.3.1. Arcs

Arcs generally represent activities which consume time and incur cost to produce performance and have a probability of successful completion. The values for these parameters are entered as constants, random variables or functions of values developed elsewhere in the network. Arcs must have both an input and an output node.

3.3.1.1. Parameters

Arcs carry both primary and cumulative values for time, cost and performance. The primary values are those established for the particular arc by means of statistical distributions or specified functions. The cumulative values for the arcs are the sums of the cumulative values for the input node and the primary values developed on the arc. In addition each arc carries a probability of successful completion for the represented activity. Thus, although the arc may be activated (or initiated), it may not be successfully completed.

3.3.1.2. Functions

Arc functional values are developed through the use of transformations as listed in table 1. These transformations are used by specifying the variables of X, Y, and Z as constants or previously derived values. The transformations may be used in series for more complicated functions.

Table 1
VERT Transformations

1.	$X*Y*Z = R$	14.	$X*(\text{LOG}_E\,(Y*Z)) = R$
2.	$(X*Y)/Z = R$	15.	$X*(\text{LOG}_{10}\,(Y*Z)) = R$
3.	$X/(Y*Z) = R$		
4.	$1/(X*Y*Z) = R$	16.	$X*(\text{SIN}(Y*Z)) = R$
5.	$X+Y+Z = R$	17.	$X*(\text{COS}(Y*Z)) = R$
6.	$X+Y-Z = R$	18.	$X*(\text{ARCTAN}(Y*Z)) = R$
7.	$X-Y-Z = R$	19.	If $X \geq Y$, then $Z = R$ If $X < Y$, then $Y = R$
8.	$-X-Y-Z = R$	20.	If $X \geq Y$, then $Y = R$ If $X < Y$, then $Z = R$
9.	$X*(Y+Z) = R$		
10.	$X*(Y-Z) = R$	21.	If $X \geq Y$, then $Z = R$ If $X < Y$, then $X = R$
11.	$X/(Y+Z) = R$	22.	If $X \geq Y$, then $X = R$ If $X < Y$, then $Z = R$
12.	$X/(Y-Z) = R$	23.	$(X*Y)+Z = R$
13.	$X*(Y)^Z = R$	24.	$(X*Y)-Z = R$
		25.	$(X/Y)+Z = R$
		26.	$(X/Y)-Z = R$

* Implies multiplication.
R indicates the resultant value.

3.3.1.3. Statistical Distributions

VERT uses fourteen statistical distributions to generate the appropriate random variables. These distributions are given in table 2. If the distributions are not satisfactory, the user can enter a histogram approximation.

Table 2

VERT Statistical Distribtuions

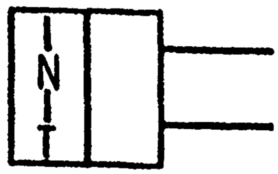
Uniform	Exponential
Triangular	Chi Square
Normal	Beta
Lognormal	Poisson
Gamma	Pascal
Weibull	Binomial
Erlang	Hypergeometric

3.3.2. Nodes

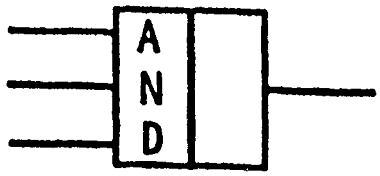
Nodes control flow in the network. The nodes are of two types--those employing split logic and those having unit logic. Split node logic employs separate logic on the input arcs into, and output arcs out of, the node. Split node logic requires that the input logic be satisfied before the output logic is initiated. Single unit logic nodes perform both input and output operations simulaneously. Nodes may have escape arcs, which are follow if the logic is not satisfied.

3.3.2.1. Split Input Logic

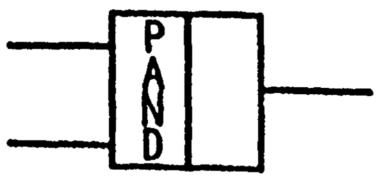
Input logic places conditions on the input arcs which must be met in order to satisfy the node requirements. If there is a chance that the input logic cannot be satisfied, an escape arc is utilized so that flow will not stop at an internal node. There are four types of split input logic.



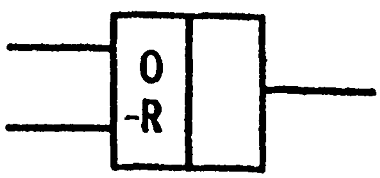
The start point of the network uses INITIAL input logic. There may be several initial nodes. Each is begun simultaneously with the same start-up parameter values.



The AND input logic requires that all input arcs be successfully completed before the output arcs are initiated.



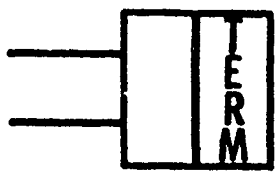
The partial and (PAND) input logic requires at least one input arc to be successfully completed but waits until all arcs have been tried.



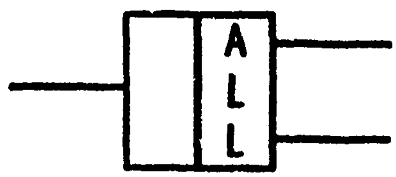
The OR input logic initiates the output logic when the first successful input arc has been completed.

3.3.2.2. Split Output Logic

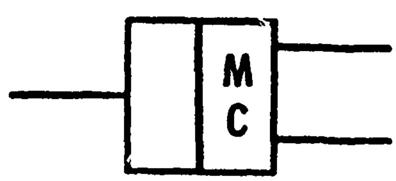
Output logic on split logic nodes is examined if the input logic has first been satisfied. Output logic determines which output arcs are to be activated. There are four types of split output logic with three of one type for a total of six.



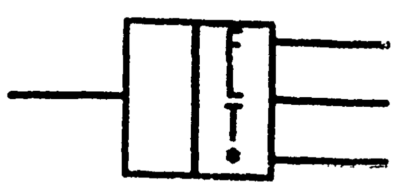
The TERMINAL node does not have any output arcs. Further explanation is given below.



The ALL logic initiates all output arcs simultaneously.



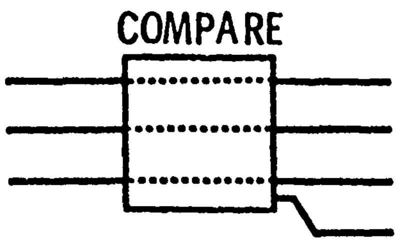
The MONTE CARLO logic initiates one and only one of the output arcs according to a user specified condition and random draw.



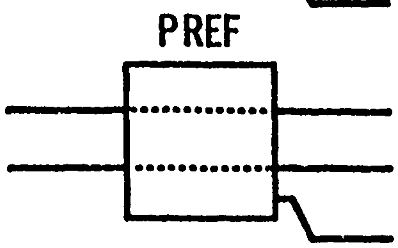
There are three types of FILTER logic which determine which arcs are to be initiated based on* (1) specified conditions on the parameters, (2) the successful completion of a specified number of input arcs, or (3) the successful completion of specified input arcs.

3.3.2.3. Unit Logic Nodes

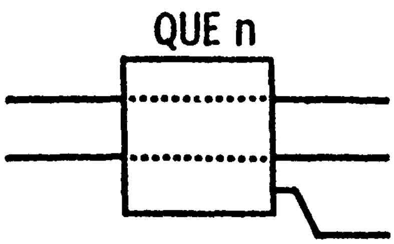
Unit logic nodes match up pairs of input and output arcs, placing conditions on the input arc to determine if the respective output arc is to be activated. The four nodes with unit logic are the COMPARE, PREFERRED, QUEUE and SORT nodes. Unit logic nodes also employ escape arcs.



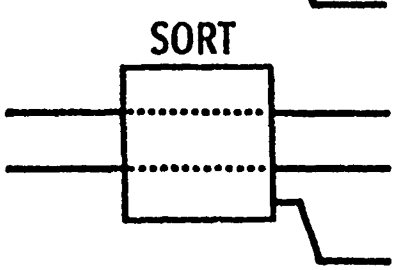
The COMPARE logic selects the optimal input-output arc set based on weighted criteria on the node parameters.



The PREFERRED logic selects the optimal input-output arc set based on user specified priorities.



The QUEUE logic processes the input-output arcs sequentially based on time of arrival and number of servers (n).



The SORT logic processes the output arcs in order of user-specified parameter weightings.

3.3.2.4. Values on Nodes

Only cumulative va. ues of time, cost and performance are computed for the interim and terminal nodes. Depending on the type of logic used by the node, the value for the node may be the sum, average, minimum or maximum of the cumulative values from the input arcs.

3.3.2.5. Terminal Nodes

Terminal nodes are used to indicate the completion of the last of a sequence of activities and to initiate the collection of statistics for the iteration. Several terminal nodes are usually used to indicate successes and failures.

3.4. Evaluation

Once the decision process has been modeled as a network using the VERT arcs and nodes, the model is evaluated through the use of a computer program designed to simulate VERT networks. The simulation is conducted by creating flow from the initial nodes and following this through to the resultant terminal nodes. One trial solution consists of tracing the flow through the network as controlled by the node logic and developing values for the arcs and nodes for evaluation of the network. The simulation process is repeated an appropriate number of times to create a distribution of possible outcomes. The computer program allows the user to select the type of output for appropriate analysis. The amount and type of output available make the program very useful in analyses.

3.4.1. Computer Program

The VERT computer program was initially written in FORTRAN IV for the IBM 360/65 computer but is currently operational on the IBM, CDC, UNIVAC and Honeywell systems. There are approximately 4000 statement cards including comments in the source deck. The core size required for a network of about 100 arcs and 100 nodes is approximately 25 k words. Execution time for 500 iterations of a network this size is less than one minute.

3.4.2. Output

The user may elect to see the value of each parameter on each arc and node for every iteration, a one-line summary of the results of each iteration, or summary results for the number of iterations. Time, cost and performance data are provided for each terminal node, a composite terminal node, requested internal nodes and intervals between internal nodes. Relative and cumulative frequency distributions are plotted for each of the

three parameter values for the terminal and selected nodes as shown in figure 3. Correlation plots and statistics as shown in figure 4 are also given for values on the terminal and requested nodes. (Because of reproduction difficulties, both figures 3 and 4 were done on a typewriter copying the computer printer output.) The critical path is indicated as well as the optimum terminal node. These are selected based on user-specified weightings applied to the parameters.

```
     Relative Frequency            Cumulative Frequency

         .05  .10  .15                 0.2  0.4  0.6  0.8  1.0
55.  I----I----I----I          55.  I----I----I----I----I----I
     I               .000           I                              .000
57.  I                         57.  I
     I               .004           I                              .004
59.  I                         59.  I
     I*              .006           I                              .010
61.  I                         61.  I
     I**             .022           I*                             .032
63.  I                         63.  I
     I******         .047           I**                            .079
65.  I                         65.  I
     I*********      .081           I****                          .160
67.  I                         67.  I
     I*************  .113           I********                      .273
69.  I                         69.  I
     I************** .124           I***********                   .397
71.  I                         71.  I
     I************** .117           I***************               .514
73.  I                         73.  I
     I************   .098           I*****************             .612
75.  I                         75.  I
     I***********    .093           I*********************         .705
77.  I                         77.  I
     I********       .065           I**********************        .770
79.  I                         79.  I
     I*******        .061           I************************      .831
81.  I                         81.  I
     I******         .052           I*************************     .883
83.  I                         83.  I
     I****           .041           I**************************    .924
85.  I                         85.  I
     I***            .032           I***************************   .956
87.  I                         87.  I
     I**             .018           I***************************   .974
89.  I                         89.  I
     I**             .015           I***************************** .989
91.  I                         91.  I
     I*              .007           I***************************** .996
93.  I                         93.  I
     I               .004           I***************************** 1.000
95.  I                         95.  I
     I----I----I----I               I----I----I----I----I----I
     0   .05  .10  .15              0   0.2  0.4  0.6  0.8  1.0
```

Figure 3

VERT Distribution Output[1]

1. The horizontal axis reflects the relative and cumulative frequency distribution of the outcome. The outcome values are reflected on the left of each chart with the respective tabular distribution reflected on the right of the chart. The parameter, for instance time, is shown as ranging from 55 to 95.

```
        xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
        y       2                              y
        y        1                             y
        y  1          1                        y
        y 1    1 1    1                        y
        y     211       1·                     y
        y     3 2 1 1 1                        y
R       yl   1 3 2 1 2 1 1                     y
E       y    21 3 1  1                         y
D       y      232234 2                        y
        y   11122332222 2                      y
P       y  1 11222333 322 11                   y
E       y      11 34672811                     y
R       y     1213 51353521 1                  y
S       y    111112355783431l 1 11 1           y
O       y 1    1 23223545241  11               y
N       y      1132244834321 211               y
N       y      1 13224769524133 1              y
E       y       1  2134656562411311            y
L       y       1  16424665 62411311           y
        y       1   34443344244151 2 11        y
L       y          3  2845*44672211 1          y
O       y             34345678215121  1        y
S       y            1  11456765431   1        y
S       y            111125566297142 121 1     y
E       y           1  1 53 5536522414         y
S       y            1 1 3134351333224 1       y
        y              112212 1454 413  1   1  y
        y              21  232413311351    21  y
        y               1 21 326332132112      y
        y                1     13414 222       y
        y                   1  2114121 1   1   y
        y                     11 2112 2    12  y
        y                    12  2   1 11  2   y
        y                       1 1211    11   y
        y                              1  1  1 y
        y                           21 1  1    y
        y                   1        2  1  2   y
        y                           1 111      y
        y                                      y
        y                                1  1  y
        y                                  1   y
        xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxy

                BLUE PERSONNEL LOSSES
```

Figure 4

VERT Correlation Output[1]

1. Numbers indicate the frequency of occurrence at each x, y point. A frequency greater than 9 is indicated by an asterisk.

4. ADAPTATION

VERT was designed to assist in the assessment of risk involved in programs and projects. However, Mr. Moeller has made the conceptual model and automated technique extremely flexible as well as powerful. The two features which make the technique so useful are (1) the use of transformations to model the three variables on the arcs as random variables or as functions of variables evaluated elsewhere in the network and (2) the amount and variety of output. The technique can be adapted to assist in the evaluation of many different complex problems, particularly force assessment. The remainder of the paper will point out how the modeling technique can be adapted for this purpose.

4.1. Use of Arc Parameters

The analyst may find that it is not possible to define all of the interim factors and pertinent measures of effectiveness (MOE) in terms of the three parameters of time, cost and performance. It is possible however to use the parameters to represent different variables on different arcs. But the pertinent MOE must appear at the terminal nodes so that output statistics can be computed in terms of them. This can be done by having the first transformation on the final arcs subtract out the cumulative values of the previous nodes. Succeeding transformations can then be made to develop the MOE from values computed earlier in the network. The values on the terminal nodes are thus made to reflect the appropriate MOE and are available for statistical analysis.

4.2. Source of Input Data

For network simulation to be useful in force analysis, it is necessary to determine meaningful relationships between the factors which affect capabilitiy. Values and relationships for interim variables used in the analysis may come from network analysis or they may be the results of separate analysis such as static assessments or computer wargame simulation. Multivariate regression on a series of wargames could provide the relationships for use in VERT in the evaluation of force mixes and other conditions not specifically wargamed.

4.3. Performance and Risk Evaluation

Network simulation provides a means to evaluate and

compare force capability. A measure of the risk can be determined which reflects the capability of a force to meet a particular mission. Figure 5 depicts a cumulative distribution reflecting what might be the output of a network simulation. The MOE represented is the number of days before a critical phase line is crossed. This MOE was chosen as an indication of the combat capability of the force when facing the threat in a particular scenario. The figure shows the distribution of results running from about 10 to 90 days. The histogram reflects the grouped results as given in the computer output. The smooth curve represents an approximation of the continuous distribution.

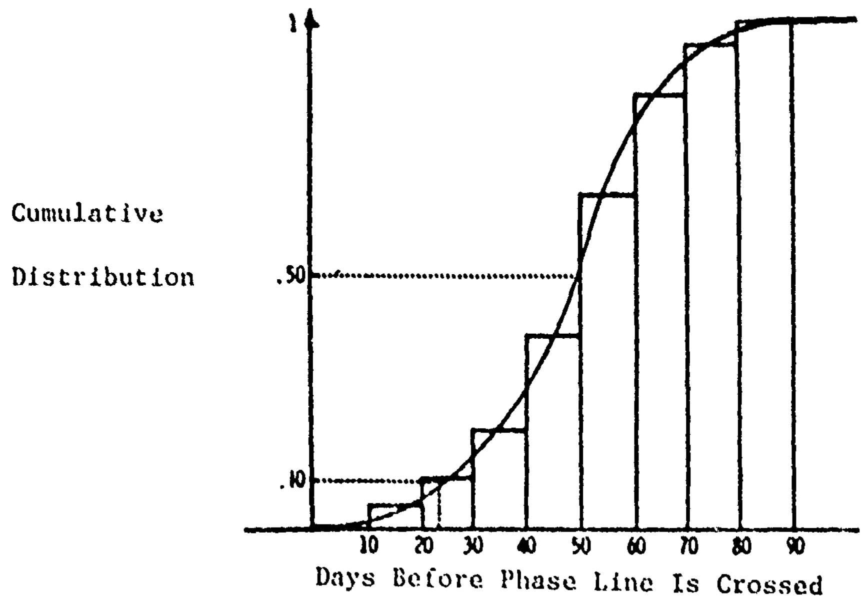


Figure 5

Example Use of Cumulative Distribution

4.3.1. Risk Analysis

Risk is defined in decision analysis as the uncertainty or chance of failure in meeting schedule, cost or performance goals. Risk can be measured in terms of the probability of not attaining a desired level of performance or meeting time and cost constraints. In this case a measure of risk may be derived by interpreting the cumulative frequency as an estimate of the probability that this force will not be able to hold the phase line at least a desired number of days. Thus by starting with the number of days which the line must be

held, say before reinforcements arrive, the level of risk associated with a particular force holding the line the required time can be determined. Conversely, the number of days can be determined for a specified risk level. For example a ten percent risk level in the above results indicates 24 days, i.e., the results indicate a ten percent risk factor for this force holding the line 24 days. Starting with a desired holding time of 50 days shows a 50 percent risk factor.

4.3.2. Comparative Performance

The cumulative distribution may be used as a measure of performance of a particular force or as a comparison measure of different forces. The results of decision rules, or the sensitivity of the results to assumptions or input variables may also be compared. Mean or median values or any of the percentage levels in the cumulative distribution may be used for comparing different force mixes or quantities. The level of risk for stated capability levels, the capability levels for stated levels of risk or combinations may be used for comparison of force capabilities.

4.4. Example Force Analysis Problem

To demonstrate the application of network simulation to the force assessment process, the following hypothetical example has been constructed. The specified relationships are stated for example only.

4.4.1. Description

The problem is to analyze the combat results when a friendly Blue force opposes an enemy Red force. The forces on both sides can be committed in two increments. The amount of forces committed on both sides in the first period is a function of the warning time, which can be treated as a random variable over a fixed range. For this example, personnel and territory loss rates are treated as functions of the force ratio. Further, as each relationship has been drawn from a regression analysis, the relationship includes variation around the respective regression line. The initial increment

of Blue forces can lose the battle if the first phase line is crossed indicating the forward edge of the battle area (FEBA) has moved too far and too much territory has been lost before reinforcements arrive. Once the reinforcements arrive, the battle can terminate in one of three ways: a break off point is reached for either Red or Blue forces because of personnel losses or the second phase line is crossed indicating again that too much territory has been lost.

4.4.2. Formulation

The network example is formulated as shown in figure 6 and table 3 indicating the relationship of arcs, activities, nodes, parameters and MOE. Activities, indicated above the arc names, may require two or more arcs because more than three variables may need to be evaluated for that activity. The explanation in the paragraphs below is given in terms of the activities.

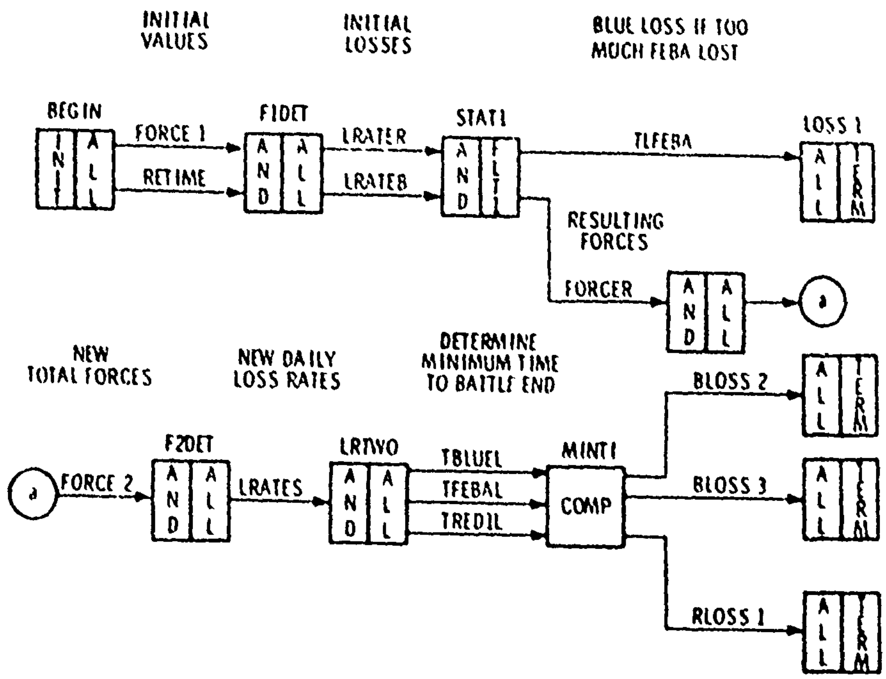


Figure 6

Example Force Analysis Network

Table 3

Arc Parameters for Example Network

Arc	"Time"	"Cost"	"Performance"
FORCE1	Warning time	Blue forces	Red forces
RETIME	-	FEBA location	Reinforcement time
LRATER	Force ratio	Red losses	
LRATEB	-	Blue losses	FEBA loss
TLFEBA	Battle time	Blue losses	FEBA loss
FORCER	FEBA location	Blue forces	Red forces
FORCE2	-	-	Force ratio
LRATES	FEBA loss rate	Red loss rate	Blue loss rate
TBLUEL	Time to Blue breakpoint	Blue losses	FEBA loss
TFEBAL	Time to FEBA breakpoint	Blue losses	FEBA loss
TRED1L	Time to Red breakpoint	Blue losses	FEBA loss
BLOSS2	Battle length	Blue losses	FEBA loss
BLOSS3	Battle length	Blue losses	FEBA loss
RLOSS1	Battle length	Blue losses	FEBA loss

4.4.2.1. Initial Values

Two arcs are used to derive the values for the following variables: warning time, reinforcement time, initial Blue and Red forces, and initial FEBA position (measured in kilometers from some zero location). Warning time is derived as a random variable with specific distribution over a fixed range, initial FEBA position is entered as a constant and the other variables were modeled as functions (the regression and residual) of the warning time.

4.4.2.2. Initial Losses

The arcs LRATER and LRATEB develop the force ratio, personnel and FEBA loss rates and compute the amounts of personnel and FEBA loss for the initial force. Loss rates are modeled as a function of the force ratio. Loss amounts are modeled as functions of loss rates, initial forces and the length of time before the reinforcements arrive.

4.4.2.3. Blue Loss If Too Much FEBA Lost

If the amount of FEBA lost exceeds a specified amount, the node STAT1 sends the flow along the arc TLFEBA and the simulation terminates. The arc TLFEBA computes the battle length, the amount of Blue personnel losses, and FEBA loss when this situation occurs.

4.4.2.4. Resulting Forces

If the initial forces sufficiently delay (without being pushed beyond the first phase line), the amount of reinforcements is determined by subtracting the initial forces from the total available Blue and Red forces. The arc FORCER determines the resultant Blue and Red forces and current FEBA location.

4.4.2.5. New Total Forces

The new force ratio is determined on the arc FORCE2.

4.4.2.6. Now Daily Loss Rates

The daily personnel and FEBA loss rates are determined on the arc LRATES, modeled as a function of the force ratio and force levels.

4.4.2.7. Determine Minimum Time to Battle End

By dividing the remaining forces and distance between the FEBA and second phase line by the daily personnel and FEBA loss rates respectively, the number of days until battle termination can be computed for each of the three possibilities of Blue loss because of personnel or territory or Red loss because of personnel. The arcs TBLUEL, TFEBAL and TRED1L compute the time and resultant losses for each case.

4.4.2.8. Termination

The node MINTI matches the input and output arcs based on the minimum time computed for the three termination possibilities. The arcs BLOSS2, BLOSS3, and RLOSS1 compute the desired MOE of battle length, total Blue personnel losses and FEBA lost for whichever path is followed.

4.4.3. Analysis

The VERT output gives the following information for analysis:

1. Frequency of termination in each terminal node. This is useful in determining probabilities of outcomes, e.g., 30% chance that initial Blue forces will not be able to survive until reinforcements arrive.

2. Distribution, statistics and correlation of outcome values for the chosen MOE of length of battle, amount of Blue force losses, and FEBA loss.

3. Distribution, statistics and correlation for the MOE on the composite terminal node indicating battle length, personnel losses and FEBA losses, irrespective of which side lost or how.

4.4.4. Utilization

By changing values, relationships, and decision logic rules, analysis can be made of the sensitivity of results to these changes.

4.5. Extensions

The example problem was a greatly simplified network. For instance, it modeled the amount of forces available in the first period only as a function of warning. Many additional factors might affect the initial force size, to include the amounts of strategic or intratheater mobility available, the availability of intermediate staging bases, and port clearance capacities in the theater. The amount of reinforcements might also be dependent on the probability of interdiction as well as the existence of rising crises in other contingency areas. These events could be modeled as extensions of the basic network. Their effect on the combat results could be modeled in accordance with the type of effect various possibilities have on the forces available and hence on the ultimate battle outcome.

Another potential extension is the inclusion of a sub-network for the air battle which would interact with the ground battle according to the amount of close air support available and its effect on the ground battle.

4.6. VERT Improvements

This paragraph describes some of the improvements made to the VERT program which have been designed to assist the analyst and make the model more responsive to the analyst's needs.

4.6.1. The VERT program requires a significant amount of the analyst's time to design and build a computer data deck for the model. Once the initial data input is prepared, the model has extensive data error checking sections which insure compatibility with the network programs. When errors are incurred, an error message is printed on the output and the processing is usually stopped, requiring corrections to the data deck. This corrected version is then rerun and the process continues until a successful completion.

4.6.2. The first improvement is the use of card-image file, created using a CRT device (terminal display screen), thus saving considerable set-up and modification time.

4.6.3. A real-time version of the batch VERT program was created to assist in the debugging phase of development. This real-time model uses the card-image file described above so that changes to the data file can be made quickly and the simulation rerun. In the production phase this real-time model also cuts down the turn around time considerably.

4.6.4. To assist in the development and modification of the network model a graphics package has been designed to provide a representation of the significant parts of the network. Three levels of detail are provided: level 1 gives a list of the arc names, respective nodes and comments (see figure 7); level 2 provides each node and its respective connecting nodes and arc names (see figure 8); level 3 draws the nodes and arcs in network display (see figure 9).

THIS IS A SUMMARY OF THE DATA FILE

ARC	NODE(FROM	- TO)	REMARKS
FORCE1	BEGIN	F1DET	DETERMINE INITIAL VALUES
RETIME	BEGIN	F1DET	DETERMINE REINF TIME 10 - W
LRATER	F1DET	STAT1	DETERMINE FORCE RATIO AND RED LOSSES
LRATEB	F1DET	STAT1	DETERMINE BLUE LOSSES AND FEBA LOSS
TLFEBA	STAT1	LOSS1	DETERMINE IF TOO MUCH FEBA LOST
FORCER	STAT1	REINF	DETERMINE TOTAL WITH REINF FORCE
FORCE2	REINF	F2DET	DETERMINE FORCE RATIO
LRATES	F2DET	LRTWO	DETERMINE LOSS RATES
TBLUEL	LRTWO	MINTI	BLUE LOSS BY PERSONNEL
TFEBAL	LRTWO	MINTI	BLUE LOSS BY FEBA LOSS
TRED1L	LRTWO	MINTI	RED LOSS BY PERSONNEL
BLOSS2	MINTI	LOSS2	BLUE LOSS DUE TO FORCES
BLOSS3	MINTI	LOSS3	BLUE LOSS DUE TO FEBA
RLOSS1	MINTI	RLOSS	RED LOSS DUE TO FORCES
XLOSS1	MINTI	DUMMY	DUMMY TO HOLD CARRY OVER

Figure 7

Data File Summary

STARTING NODE	END NODE	ARC NAME
BEGIN	F1DET	FORCE1
	F1DET	RETIME
F1DET	STAT1	LRATER
	STAT1	LRATED
STAT1	LOSS1	TLFEDA
	REINF	FORCER
REINF	F2DET	FORCE2
F2DET	LRTWO	LRATES
LRTWO	RINTI	TBLUEL
	RINTI	TFEDAL
	RINTI	TREDIL
RINTI	LOSS2	BLOSS2
	LOSS3	BLOSS3
	RLOSS	RLOSS1
	DUMMY	XLOSS1

Figure 8

Arc Node Relationship

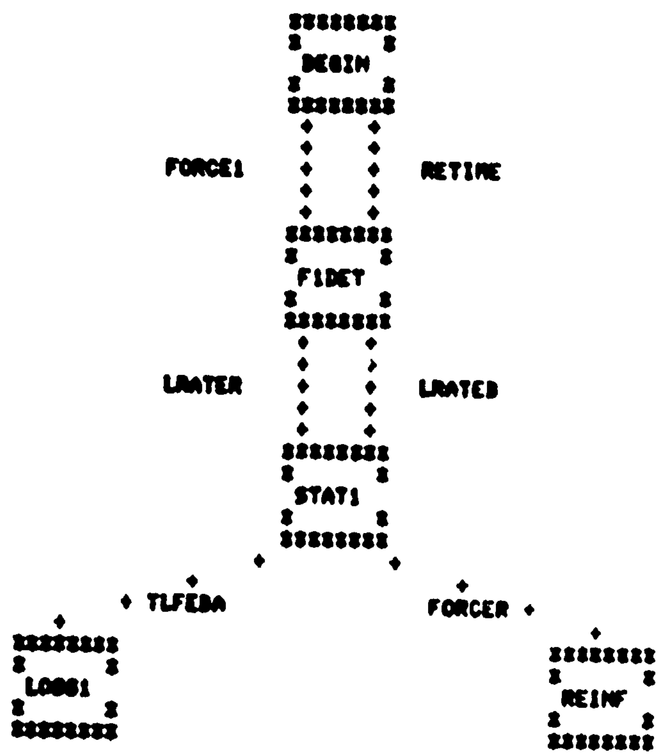


Figure 9

Network Flowchart

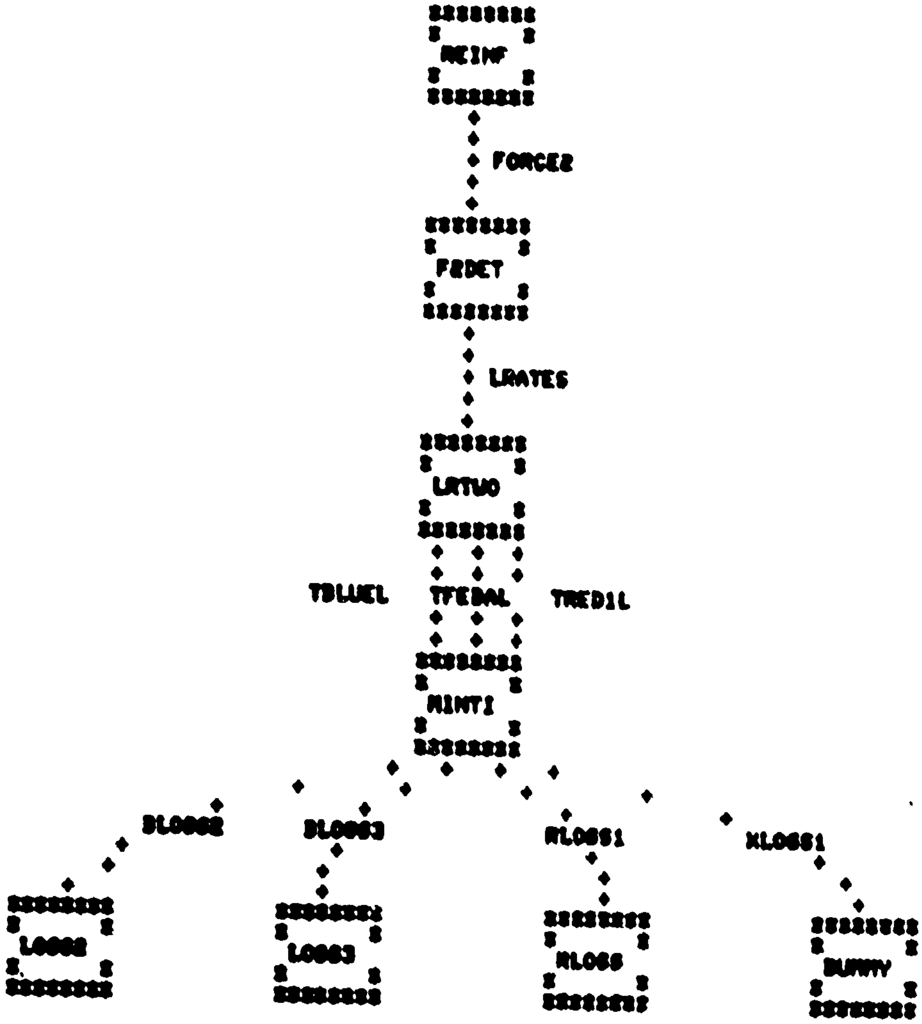


Figure 9 (cont)

5. UTILITY OF NETWORK SIMULATION IN FORCE ANALYSIS

The use of network simulation in an application such as the hypothetical example or an extension of it would highlight the activities, assumptions, and relationships to which the outcome of each alternative is sensitive. For instance, analysis might show that the battle outcome is more sensitive to the time distribution of the arrival of Blue forces than to the total amount. This would indicate improvements in mobility or forward basing would add more to the chances of success than increases in the force size.

Another valuable use is the investigation of the criticality of assumptions. If, for example, the chances of battle success are good when ammunition for the force is avaiable in great amounts, but are considerably decreased when the ammunition available is limited, then additional funds for ammunitions may produce better results than additional forces.

For the model to be effective and useful, the analyst using network simulation must assure that the model realistically portrays the relationships which exist. The entire problem needs to be correctly defined with the proper alternatives and interrelationships of events. It is necessary then that the analyst be completely familiar with not only the entire force development process, but also the decisionmakers' perspectives and preferences in the process.

6. SUMMARY

Because of the many facets of force assessment, there exists a need for a methodology to synthesize the results of many separate analyses. Currently no methodology or technique does this adequately. There is also a need to incorporate the variability of input factors and assumptions and thus provide a measure of uncertainty regarding the output measures of effectiveness.

Network simulation, particularly using VERT, is very flexible and can be adapted to incorporate results from

various analyses. VERT allows the user to investigate the effects of assumptions, input data and decision rules on force capability or other measures used in force assessment.

The application of network simulation is a new approach to force assessment. It shows considerable potential to handle many parts of the problem previously only partially dealt with.

COMBAT EFFECTIVENESS, COEAs AND
TRAINING ANALYSIS

LINDSAY F. PHILLIPS

US Army TRADOC Systems Analysis Activity
White Sands Missile Range, NM 88002, USA

ABSTRACT. A key problem in the conducting of a COEA study is the analysis of the Operational Effectiveness of a given Weapon System in relation to the Combat Effectiveness of a Combined Arms Combat Team. Similarly, a knowledge of the required minimum Combat Effectiveness of the Combined Arms Combat Team and the required minimum Operational Effectiveness of a given Weapon System organic to that team leads to a set of minimum performance standards which may be used as the basis for developing new training programs or as a tool in analyzing present training programs.

This paper outlines a procedure, using Force-on-Force combat simulation models, which relates a desired level of Combat Effectiveness to performance parameters of a given Weapon System. A methodology is introduced utilizing the results of these procedures in a COEA analysis. The application of these procedures to training analysis is discussed. Examples are given. The output of COEA studies and training analysis using these methodologies will lead, for COEAs, to weapons systems ranked in terms of their probability of meeting combat goals, design goals, and cost goals in that order, and for training analysis studies, to minimum training goals which will permit the achievement of the desired combat effectiveness goals.

1. INTRODUCTION

The US Army (USA) conducts a Cost and Operational Effectiveness Analysis (COEA) study of each major Weapon System (WS) to be added to the USA weapons inventory. A key problem in conducting a COEA study is the analysis of the Operational Effectiveness (OE) of a WS when employed in combat within a Combined Arms Combat Team (CACT) usually of reinforced company size or larger. In addition, the training subsystem of a major WS is analyzed through a Cost and Training Effectiveness Analysis (CTEA) study. In many cases, the training and performance standards developed in a training package may have little relation to the minimum OE requirements of a WS used with a CACT in a tactical situation. This paper will present a methodology, using Force-on-Force (FOF) combat simulation models which will allow the analyst to estimate the minimum OE requirements of a WS in a CACT in terms of Combat Effectiveness (CE) parameters, relate these minimum requirements to training and performance standards, and finally to rank order a weapon system in terms of its ability to meet combat requirements, design specifications, and cost limitations in a form easily digestible to decision makers.

2. METHODOLOGY

2.1 Basic Ideas

2.1.1 Weapon System Performance Parameters

Given a particular combat scenario, the OE of most infantry type weapon systems will depend upon the Probability of Hit (P_H), Probability of Kill (P_K), and Rate of Fire (R_F) of the weapon assuming man-in-the-loop. Other WS parameters which effect OE and which involve such parameters as Reliability and Maintainability (RAM), Logistics, etc., can be reduced to perturbations or P_H or R_F. The P_K, given a hit, is an empirically determined parameter dependent only upon the physical characteristics of the warhead and the target and independent of P_H and R_F. Hence, it is assumed that the fundamental WS parameters of interest, which also include the effects of the man-machine interface, are simply the P_H and R_F of the weapon system.

2.1.2 Minimum Required Battle Outcome

It is often assumed or "required" that the "friendly" forces must win the first battle of the next war. This is an ideal and not a minimum acceptable requirement suitable for analysis. For purposes of analysis, it is more meaningful, both militarily and analytically, to require, as a minimum requirement of the outcome of a battle, that the friendly forces, i.e., the CACT under consideration, shall not lose the first battle of the next war or any other battle for that matter. This requirement allows a stalemated battle condition, i.e., a battle in which neither side "wins". This idea, coupled with the idea of P_H and R_F describing the essential combat characteristics of a weapon system, including the effects of man, will allow the analyst, through the use of combat simulation models, to determine the weapon system parameters, P_H and R_F, which will produce measures of the minimum required combat effectiveness of a CACT in a stalemated battle under given scenario conditions. The minimum required combat effectiveness will be expressed in terms of the initial force ratio (F_R) defined as the ratio of unfriendly forces to friendly forces.

2.2 Procedure

2.2.1 Use of Force-on-Force Models

The use of large scale, Force-on-Force computer simulation models is required. The model may be either stochastic or deterministic. We have found that deterministic models, based upon the application of Lanchester's equations of war, are satisfactory and require much less labor and time than the stochastic models to obtain a useful result.

To date, we have applied the TRASANA combat simulation model using a European scenario in the analysis of training standards for two Anti-Tank Guided Missile (ATGM) systems. The P_H and R_F required to produce a stalemated battle condition against an opponent of initial force ratio (F_R) in a given scenario were obtained rather quickly with minimum effort assuming no requirement to completely design the basic combat scenario model including the terrain model to be used. The results of these studies were instrumental in influencing the weapon procurement decision making process.

2.2.2 Stalemated Battle Outcome Condition

The Fractional Exchange Ratio, (FER), defined as

$$FER = \frac{\% R_L}{\% B_L} \qquad (1)$$

where $\% R_L$ is the percent of RED losses (% unfriendly force losses) and $\% B_L$ is the percent of BLUE losses (% friendly force losses), is often used to determine battle outcome. We have used an end-of-battle condition to be that point at which either RED or BLUE forces have suffered an attrition of 60 percent. At this point the FER is calculated and if

1. FER > 1, then BLUE is considered to be winning the battle since the force ratio has decreased during the battle,

2. FER < 1, then BLUE is considered to be losing the battle, since the force ratio has increased, and if

3. FER = 1, then the force ratio has remained unchanged and the battle is considered a stalemate with neither side "winning". We take the condition of

$$FER = 1.0 \qquad (2)$$

to be the required minimum acceptable battle outcome, namely that of a stalemated battle.

2.2.3 Development of the FER = 1.0 Envelope

Our task is to develop an envelope of P_H, R_F values vs. F_R under the conditions of FER = 1.0. This is accomplished as follows:

1. Develop a set of scenarios, say k in number, differing only in the size and composition of the RED forces. This will provide k scenarios differing only in the F_R. This is easy to accomplish if a "standard scenario" is available which can be changed by adding or deleting units of the RED forces by simply changing certain scenario input parameters.

2. Arbitrarily select a set of P_H and R_F parameters, say m and n in number respectively. This will provide m values of P_H and n values of R_F.

3. Choose a value of P_H and R_F, denoted by P_H^M and R_F^N respectively, where the capital superscripts denote particular values of P_H and R_F, and, while holding these values constant for each simulation, make a combat simulation run for each of the k values of F_R where F_R is the force ratio which label each of the scenarios to be used. Compute the FER at the end of each battle.

4. Plot the FER vs F_R curve. In general, the results will resemble Figure 1.

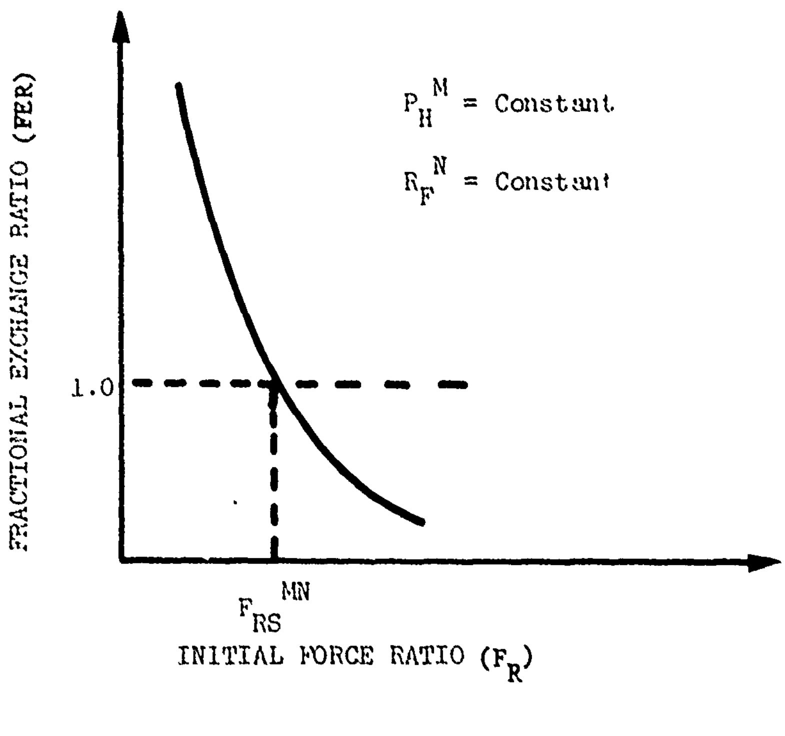


Fig. 1 Fractional Exchange Ratio vs. Initial Force Ratio

The functional form of the FER vs. F_R curve can be estimated to any degree of accuracy by performing a polynomial fit in the Least Squares sense. We have not found it necessary to perform a fit of any higher degree than that of a linear fit.

5. Calculate the value of F_R at the intersection of the FER = 1 and FER vs F_R curves. Denote this value of F_R at FER = 1.0 by F_{RS}^{MN} which corresponds to a value of F_R which can be stalemated for the constant values of P_H^M and R_F^N used.

6. Repeat steps 4. and 5. for all values of P_H and plot the P_H vs F_{RS}^{mN} curve where N denotes the constant value of R_F^N used and m denotes the set of the values of P_H^m used. The result will resemble that of Figure 2 for FER = 1.0.

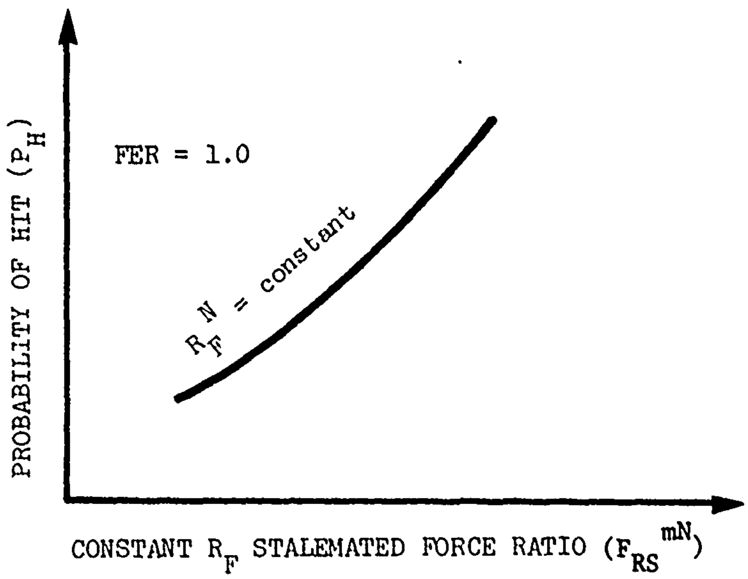


Fig. 2 P_H vs. F_{RS}^{mN} Curve at FER = 1.0

The functional form of the P_H vs. F_{RS}^{mN} can often be satisfactorily estimated by a Least Square linear fit to the data.

7. Steps 4. and 5. and 6., are repeated for each value of R_F choosen. Upon completion, one obtains an envelope of (P_H, R_F, F_{RS}) values as shown in Figure 3.

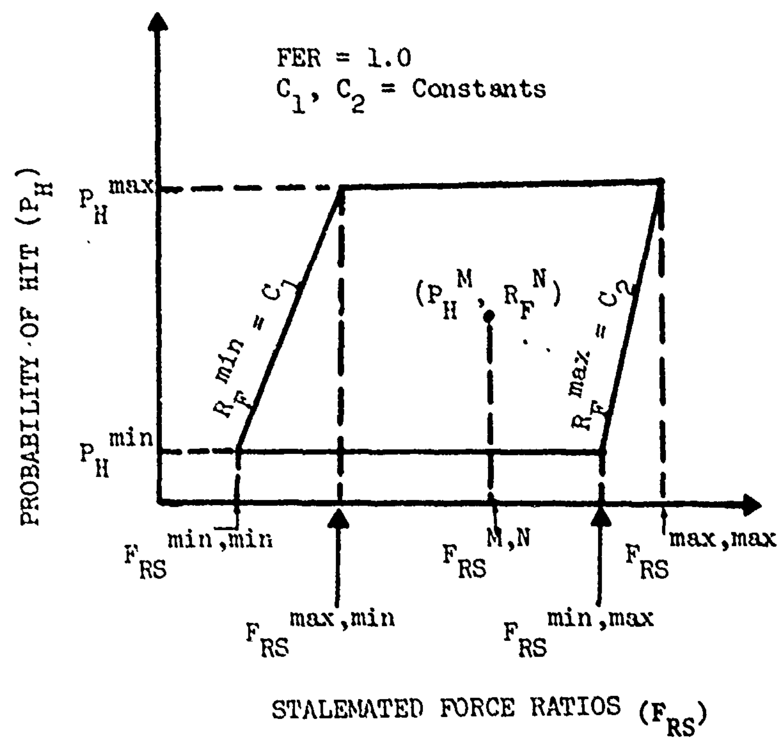


Fig. 3 The (P_H, R_F, F_{RS}) Envelope

The envelope is bounded by two Iso — R_F lines corresponding to the minimum and maximum values of R_F chosen and by the lines $P_H = P_H^{min}$ and $P_H = P_H^{max}$ corresponding to the

extreme values of P_H chosen. Any particular (P_H^M, R_F^N) pair within the envelope has associated with it a value of the force ratio which may be stalemated, F_{RS}^{MN}, with these values of P_H and R_F.

The total number of computer simulation runs required to generate Figure 3 will be no greater than the cartesian product of the maximum values of m, n, and k. We have obtained useful results with as few as 18 computer runs corresponding to the maximum choosen values of (m, n, k) being (3, 2, 3).

3. APPLICATIONS OF THE FER = 1.0 ENVELOPE

3.1 Cost and Operational Effectiveness Analysis (COEA)

3.1.1 Interpretation of FER = 1.0 Envelope

Weapon Systems are designed with a P_H and R_F specified. Suppose this design point corresponds to the point $D(P_{HD}, R_{FD})$ in Figure 4.

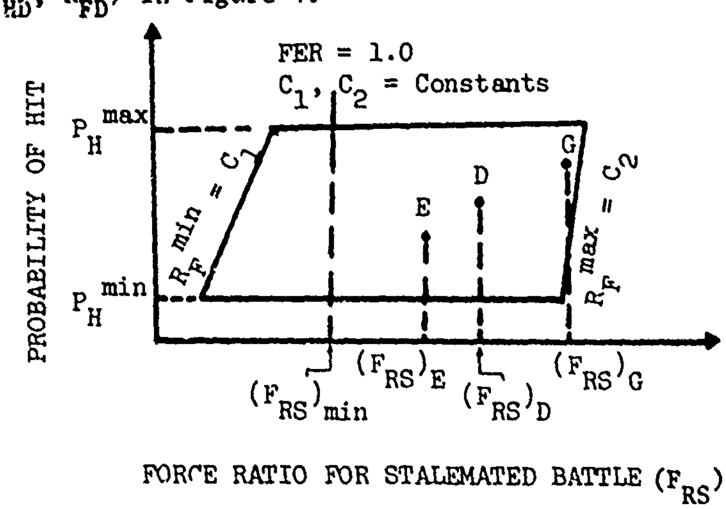


Fig 4 Performance Requirements Overlaid on FER = 1.0

Suppose further that it is required to achieve a stalemated battle condition corresponding to FER = 1.0 against some initially unfavorable force ratio of $F_R > 1$ no smaller than that designated by $(F_{RS})_{min}$. That portion of the (P_H, R_F, F_{RS}) envelope to the right of the line $F_{RS} = (F_{RS})_{min}$ represents an excess of weapon system combat effectiveness capability to meet the minimum required threat successfully. That portion to the left represents a deficiency in weapon system capability to meet the required minimum threats, subject to the constraints of the scenario including, but not limited to, tactics, weapons mix, terrain, etc.

3.1.2 Probability of Achieving Weapon System Performance Parameters

Associated with each general weapon system parameter pair $G(P_{HG}, R_{FG})$, there exists a probability P(G) that the WS will achieve that performance. Let $D(P_{HD}, R_{FD})$ be the parameter pair point corresponding to the design parameters of the WS and let P(D) denote the probability of WS achieving that design point. Physically, P(G) should be a bivariate distribution centered, most generally, in an asymmetrical fashion about the design point D, with P(G) decreasing more rapidly in the direction of increasing P_H and R_F and decreasing more slowly as P_H and R_F decrease. A symmetrical distribution is possible only in very special cases. For example, $P_H = .5$ could lead to a symmetrical bivariate distribution with R_F. Preliminary calculations could assume any physically reasonable distribution. Field test data will provide additional information regarding the distribution P(G). The expected value $E(P_{HE}, R_{FE})$ belonging to the set $G(P_{HG}, R_{FG})$ can then be calculated. The points $G(P_{HG}, R_{FG})$, $D(P_{HD}, R_{FD})$, and $E(P_{HE}, R_{FE})$ are shown in Figure 4. The line $F_{RS} = (F_{RS})_{min}$ and the value of F_{RS} corresponding to the point G are also shown.

3.1.3 Rank Ordering

Suppose there are "i" weapon systems which must be compared. Let "b" denote the base system, usually the currently fielded system, to which all other systems are to be compared. Given the values of $(F_{RS})_E$, $(F_{RS})_D$ and $(F_{RS})_{min}$ corresponding to the stalemated force ratio which can be achieved by the expected values (P_{HE}, R_{FE}), the stalemated force ratio which can be achieved by the design parameters (P_{HD}, R_{FD}), and the minimum force ratio which must be stalemated respectively, one performs the following calculations where the superscript or subscript "i" denotes quantities associated with the ith weapon system under consideration and the superscript or subscript "b" refers to the base system:

Let $\quad \Delta K_i = K_i - K_b \qquad (3)$

where $\quad K_i = (F_{RS})_E{}^i - (F_{RS})_{min}$

and $\quad K_b = (F_{RS})_E{}^b - (F_{RS})_{min}$

Also let $\quad \Delta D_i = D_i - D_b \qquad (4)$

where $\quad D_i = (F_{RS})_E{}^i - (F_{RS})_D{}^b$

and $\quad D_b = (F_{RS})_E{}^b - (F_{RS})_D{}^b$

The difference, ΔK_i, is a measure of how well the ith system is expected to meet its combat effectiveness requirement relative to the base case. The quantity, ΔD_i, is a

measure of how well the ith system can meet its design requirements relative to the base case. The system with the highest ΔK_i will be the most combat effective system. The maximum ΔD_i will identify the most design effective system. If one now introduces a cost factor $\Delta \$_i$ defined by

$$\Delta \$_i = COST_b - COST_i \tag{5}$$

where $COST_b$ is the cost of producing one unit of the base case system and $COST_i$ is the cost of producing one unit of the ith system, then a simple rank ordering of all the systems in descending order of ΔK_i, ΔD_i, and $\Delta \$_i$ will produce a rank ordering of the weapon systems under consideration in terms of combat/operational effectiveness, design effectiveness, and cost effectiveness respectively The weights assigned to each factor, ΔK_i, ΔD_i, and $\Delta \$_i$, are left to the discretion of the decision maker since the arbitrary combination of these three factors will produce a single arbitrary rank ordering, essentially dependent upon the weight system used.

3.2 Training Applications

3.2.1 Determination of Minimum Training Standards

From the previous discussion, it is known what values of P_H and R_F lie along the line $F_{RS} = (F_{RS})_{min}$. As long as trainees can perform with the weapon system such that their P_H and R_F parameters lie to the right of the $(F_{RS})_{min}$ line, then the weapon system, including man, will meet the minimum combat effectiveness requirements. A weapon system unit is technically "combat ready" if the crew of that unit can produce a performance to the right of the $(F_{RS})_{min}$ line of Figure 4.

One possibility of determining minimum training standards is to define some arbitrary increment δ of F_{RS}. Those trainees performing to the left of $[(F_{RS})_{min} - \delta]$ line as indicated in Figure 5, would be terminated in their training

due to unsatisfactory performance whereas those performing better than $[(F_{RS})_{min} + \delta]$ would satisfy all training performance requirements and be available for the combat units. Those performing in the transition range of $[(F_{RS})_{min} \pm \delta]$ could be either recycled or terminated from training if their performance did not meet standards. Although the increment δ is somewhat arbitrary, a reasonable value could be estimated from the FER = 1.0 envelope and from weapon system performance data or by assuming some reasonable lower limits to P_H and R_F based on weapon system design specifications. The percentage of trainees who are able to meet certain P_H, R_F criteria might also be of assistance in determining a value of δ. However when δ is determined, it must be translated into minimum acceptable values of the performance parameters.

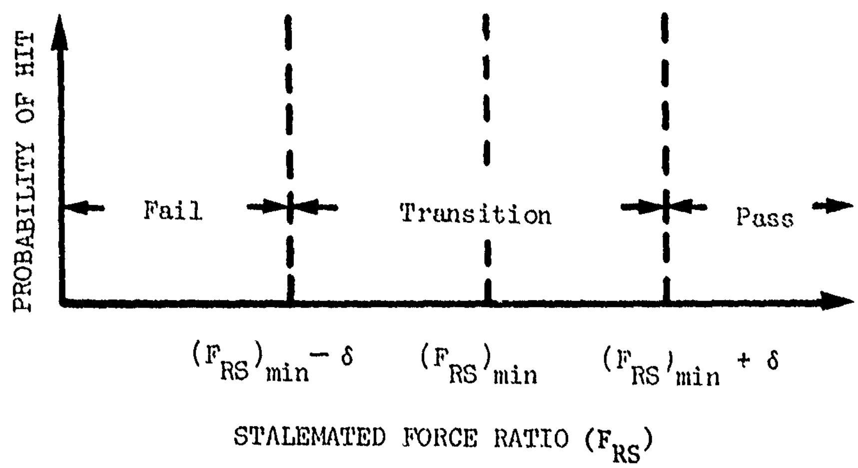


Fig. 5 Training Criteria in terms of $(F_{RS})_{min}$

3.2.2 Revision of Training Standards

Suppose their exists a fielded weapon system and training program. Suppose further that the training program is based upon a set of standards determined by "experience" with the fielded system and "knowledge" of trainee requirements and not upon analytically derived weapon system combat effectiveness requirements. Quite often the product of such a training program fails to meet the quality expectations of the training program developers. This often drives a training improvement program, sometimes of massive proportions and impact on training resources. By performing an analysis of the minimum acceptable weapon system performance parameter values by the methods proposed in this paper, i.e., by analyzing the FER = 1.0 envelope, one can readily develop revised training standards which are compatible with the combat effectiveness requirements. It is our experience that current training requirements are often more stringent than necessary. Our analysis of two ATGM systems led to cost savings in both training programs without sacrificing combat efficiency.

4. FINAL REMARKS

4.1 Usefulness of Methodology

The development and use of an envelope of weapon system performance parameters corresponding to a parameter related to stalemated battle conditions has proved itself to be a simple, quick, inexpensive, and extremely valuable tool in analyzing weapons performance requirements and training requirements in terms of overall unit combat effectiveness. Of course, the procedures we have followed could have been performed for any given battle outcome as defined by any constant FER. The choice of FER is the choice of the analyst and the problem which is being addressed.

4.2 Applicability of Methodology

In general, our methodology is applicable to any three parameter problem under the condition of a fourth parameter remaining constant. All that is required is an appropriate computer model which can generate the relationship between the four parameters. It is our feeling that a wide variety of problems can be successfully attacked with the methodology we have described.

4.3 Accuracy of Methodology

The accuracy of the method is dependent upon the characteristics of the computer model used and the method of fitting analytical curves to the data generated. In our work, we were satisfied that reasonably elementary data fitting methods were consistant with model accuracies and our analytical objectives. If one wishes a "ball park" estimate, then useful results can be obtained in days rather than weeks or even months. Herein lies the value of our methodology for related combat effectiveness of a combat unit to the performance parameters of a given weapon system employed by that combat unit.

RESOURCE ALLOCATIONS IN FORCE-ON-FORCE/FORCE MIX ANALYSIS

ROBERT P. HUNT, PHILLIP N. BILLINGSLEY, R. CARROLL LAIL

Special Projects Division,
Resource Analysis Branch,
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, NM 88002, U.S.A.

ABSTRACT. This paper presents the models used at TRASANA in the evaluation of alternative force resource requirements. These models fulfill the amplified need to relate and quantify cost and combat effectiveness and assess the impact on out year defense planning for a new weapon system. Using current COEA and study efforts, this paper presents and discusses the resource analysis results and implications using the Force Stratification System (FSS) and RECAP models. The paper includes the analytic workings of the models, as well as a discussion of the analysis results of the model output. As a summary, the paper presents weaknesses in the current force cost methodology and a projection to future force cost trends and requirements.

1. INTRODUCTION

1.1. In today's environment, changing economic values and rapid advances in technology make both cost and effectiveness central issues in the acquisition of a major weapon system. This paper will present and discuss those resource (cost) models used at the Training and Doctrine Command Systems Analysis Activity (TRASANA) in the evaluation of alternative force resource requirements in the context of an Army Cost and Operational Effectiveness Analysis (COEA). The models, which display resource requirements, are the Force Stratification system and the Resource/Cost Analysis programs.

1.2. The display of resource requirements associated with various force alternatives serves two basic purposes. This display of resource requirements allows the analyst to evaluate near and out-year funding requirements which effect various budgetary decisions, i.e, how much procurement money is required or what is the impact on personnel requirements. Also, a display of alternative force costs allows the analyst to calculate a relative cost factor for each alternative. (The relative cost of an alternative is defined as the cost for an alternative over a specific period of time, normally 20 years, divided by the cost of the base case over the same period of time.) The relative cost factor serves as an index with which cost and effectiveness can be directly compared.

1.3. TRASANA's force resource models are computer programs which sort, classify, and redistribute the forces resources defined by using several data files in the systems library.

1.3.1. One data file contains coded information on personnel, equipment, and units, It contains functional coding for all special skill identifiers (SSI) and military occupational speciality (MOS), civilian job specialities, and all TOE (Table of Organization and Equipment) items of equipment. It also contains functional and operational characteristic coding for all TOE units.

1.3.2. The other files contain the most current data available and are obtained directly from other Army force planning and accounting processes. These files contain the master authorization data required to determine the quantity of resources assigned or authorized to a TOE. They also contain information on the dollar cost of personnel and equipment. The information on TOE units is obtained from the TOE Master File maintained by the US Army TRADOC. Cost information, which has been approved by the Comptroller of the Army, is obtained from the Force Cost Information System (FCIS).

These cost data exist in a nonrecurring and annual recurring format.

2. FORCE STRATIFICATION SYSTEM

2.2. The Force Stratification System (FSS) is really two separate force analysis processes. The larger process is designed to analyze TOE forces by Standard Requirements Code (SRC). The smaller process is designed primarily to analyze Non-TOE forces. Each process has application to both current and future force analysis.

2.1.1. FSS provides an additional insight into an Army force. It allows the military planner to look beyond branches or units of the force and view all resources. FSS displays those resources that have been assigned to a particular functional area and accounts for the number of personnel and the nonrecurring and annual recurring dollar cost of personnel and equipment assigned to each function.

2.1.2. FSS is a multiphase sorting and classification process. The first phase deals with sorting and classifying personnel and equipment based on the functions they perform. The 64 functional areas into which resources are stratified are shown in Figure 1.

I. DESTRUCTION OF THE ENEMY
- 1 INFANTRY
- 2 ARMOR
- 3 CAVALRY
- 4 MORTAR ROCKET & ARTY SPT
- 5 MISSILE FIRE SUPPORT
- 6 AERIAL FIRE SUPPORT
- 7 AIR DEFENSE
- 8 GENERALIST IN DESTRUCTION OF THE ENEMY

II. INTELLIGENCE
- 9 COMBAT SURVEILLANCE & TARGET ACQUISITION
- 10 MILITARY INTEL ACTIVITIES
- 11 DENIAL, DECEPTION & COUNTERINTELLIGENCE
- 12 TERRAIN & TOPOGRAPHIC
- 13 GENERALIST IN INTELLIGENCE

III. MOVEMENT
- 14 TACTICAL AIRMOBILITY
- 15 TACTICAL GROUND MOBILITY
- 16 TACTICAL WATER MOBILITY
- 17 AIR MOVEMENT
- 18 GROUND MOVEMENT
- 19 WATER MOVEMENT
- 20 RAIL MOVEMENT
- 21 TERMINAL OPERATIONS
- 22 MOVEMENT CONTROL
- 23 GENERALIST IN MOVEMENT

IV. COMMUNICATION
- 24 RADIO COMMUNICATION
- 25 WIRE COMMUNICATION
- 26 OTHER COMMUNICATION
- 27 COMM CONST SUPPORT
- 28 GENERALIST IN COMM

V. SUPPLY
- 29 SUPPLY, GENERAL
- 30 AMMUNITION SUPPLY
- 31 POL SUPPLY
- 32 GENERALIST IN SUPPLY

VI. MAINTENANCE
- 33 AIRCRAFT MAINTENANCE
- 34 GROUND VEHICLE MAINT
- 35 WATER CRAFT MAINT
- 36 RAIL EQUIPMENT MAINT
- 37 COMM EQUIPMENT MAINT
- 38 WEAPONS MAINTENANCE
- 39 WEAPONS ELECTRONICS MAINT
- 40 OTHER ELECTRONIC MAINT
- 41 REAL PROPERTY MAINTENANCE
- 42 OTHER TYPE MAINTENANCE
- 43 GENERALIST IN MAINTENANCE

VII. ENGINEER
- 44 COMBAT ENGINEERING
- 45 CONSTRUCTION ENGINEERING
- 46 ENGINEER SPT SERVICES
- 47 GENERALIST IN ENGINEERING

VIII. MEDICAL
- 48 PREVENTIVE MEDICINE, EMER TREATMENT, EVACUATION & RESCUE
- 49 MEDICAL TREATMENT
- 50 MEDICAL LABORATORY SERVICE
- 51 DENTAL SERVICE
- 52 VETERINARY SERVICE

IX. SERVICES
- 53 LEGAL SERVICE
- 54 FINANCIAL SERVICE
- 55 ADMIN & PERSONNEL SERVICE
- 56 CHEMICAL SERVICE
- 57 LAW ENFORCEMENT
- 58 CIVIL AFFAIRS/PSYOP
- 59 FOOD SERVICE
- 60 DATA PROCESSING
- 61 OTHER SERVICES
- 62 GENERALIST IN SERVICE
- 63 LOGISTIC GENERALIST

X. MISCELLANEOUS
- 64 MISCELLANEOUS

Figure 1 Force Stratification Functions

The functional areas range from combat functions to combat service support functions. These functions are unique to FSS and represent a detailed look at resource allocations. For ease of discussion, the 64 functions are also "rolled-up" into 10 summary functions.

2.1.3. The second, third, and forth phases of the multiphase sorting and classification process deal with the operational characteristics of a resource. As the multiphase sorting process proceeds, a force stratification matrix, Figure 2, of the resources is developed.

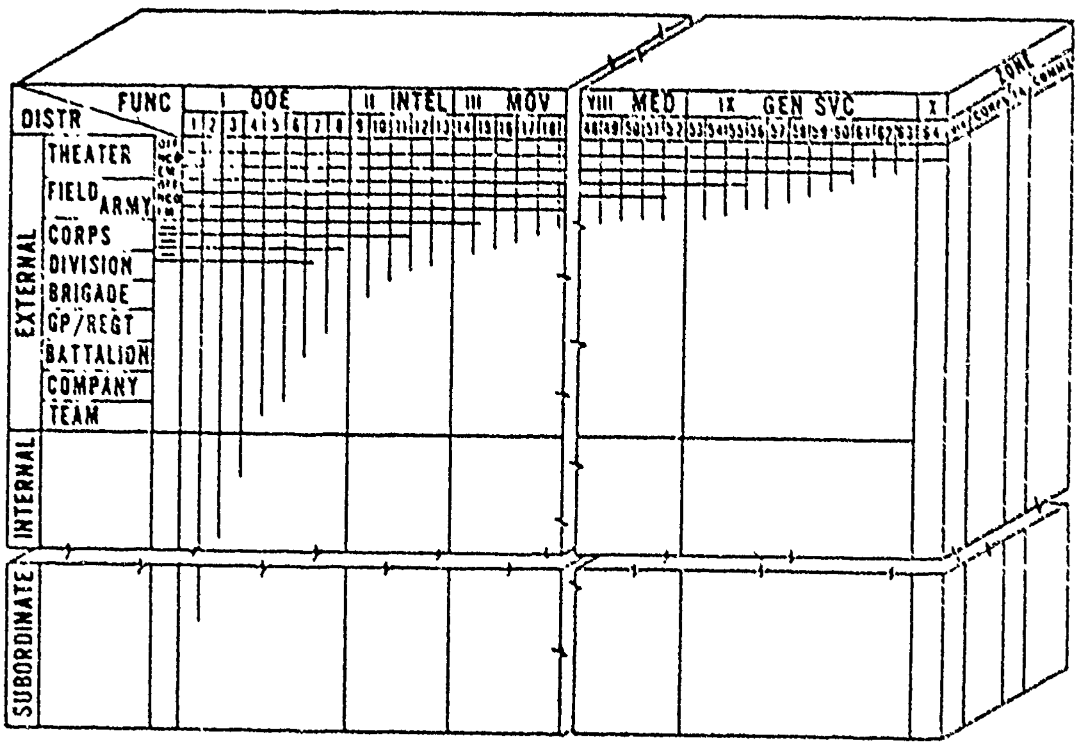


Figure 2 Force Stratification Matrix

The following are examples of other information that is available:

1. Beneficiary (See Figure 6 for definitions)
   (i) external,
   (ii) internal,
   (iii) subordinate.
2. Command Level
   (i) theater,
   (ii) field Army
   (iii) corps,
   (iv) division
   (v) brigade,
   (vi) group/regiment,
   (vii) battalion,
   (viii) company,
   (ix) team.
3. Other
   (i) branch/area,
   (ii) centralization/decentralization,
   (iii) span of control.

2.2. Examples of FSS Output

2.2.1. If an analyst were interested in knowing what percent of the personnel in a certain force were used for medical support, he might follow the plan presented in Figure 3.

COMBAT	STRENGTH
INFANTRY	100,000
ARMOR	60,000
••	••
TOTAL COMBAT	240,000
COMBAT SUPPORT	
AVIATION	10,000
TRANSPORTATION	35,000
••	••
TOTAL COMBAT SUPPORT	110,000
COMBAT SERVICE SUPPORT	
ADJUTANT GENERAL	7,000
MEDICAL	40,000
••	••
TOTAL COMBAT SERVICE SUPPORT	150,000
	TOTAL 500,000

TYPICAL ANALYSIS:
$$\text{MEDICAL SUPPORT} = \frac{40,000}{500,000} = 8\%$$

Figure 3 US Army Troop List

In this analysis, the total number of personnel is calculated in the force and the total number of personnel in medical units is calculated to form a ratio, 8 percent in Figure 3. This approach, commonly called the TOE method, has at least three disadvantages. These disadvantages are presented in Figure 4.

DISADVANTAGES OF FORMER SYSTEM

- MANY PERSONNEL IN MEDICAL UNITS ARE NOT PERFORMING MEDICAL FUNCTIONS, FOR EXAMPLE, CLERKS, MAINTENANCE, SUPPLY
- MANY MEDICAL PERSONNEL IN NON-MEDICAL UNITS, FOR EXAMPLE, MEDICS IN AN INFANTRY BATTALION
- MEDICAL COSTS ARE NOT REPRESENTATIVE OF THE ACTUAL MEDICAL COSTS IN THE FORCE

Figure 4 Disadvantages of Former System

Figure 5 presents a comparison of the results obtainable from the TOE method and FSS or MOS method. . This comparison shows different results are obtained for each method in the medical support.

MANPOWER DISTRIBUTION - HEAVY AND LIGHT CORPS

TABLE OF ORGANIZATION & EQUIPMENT (TOE) vs MILITARY OCCUPATIONAL SPECIALTY (MOS)

(PERCENT OF PERSONNEL DEVOTED TO FUNCTION)

SUMMARY FUNCTIONS	HEAVY CORPS TOE	HEAVY CORPS MOS	LIGHT CORPS TOE	LIGHT CORPS MOS
I DESTRUCTION OF ENEMY	48.9	27.0	51.0	29.7
II INTELLIGENCE	2.9	3.2	2.9	3.1
III MOVEMENT	8.0	8.1	8.3	8.7
IV COMMUNICATION	4.9	7.9	4.8	7.4
V SUPPLY	2.0	8.3	2.6	8.5
VI MAINTENANCE	14.2	19.8	12.2	18.1
VII ENGINEERING	5.0	6.4	7.3	5.2
VIII MEDICAL	8.1	5.6	6.1	5.6
IX SERVICES	5.0	13.6	5.0	13.6
X MISCELLANEOUS	0.0	0.1	0.0	0.1

Figure 5 Manpower Distribution

In addition to the comparison of medical limits, Figure 5 presents the difference obtained from the various methods at the ten summary levels.

2.2.2. FSS may be used to present beneficiary stratification of various units. Figure 6 iefines beneficiaries of functions.

BENEFICIARIES OF FUNCTIONS

- ♦ EXTERNAL FUNCTION - ONE WHICH IS BEING PERFORMED FOR THE TOTAL FORCE
- ♦ INTERNAL FUNCTION - ONE WHICH IS BEING PERFORMED ONLY FOR THE UNIT TO WHICH THE RESOURCE IS ASSIGNED
- ♦ SUBORDINATE FUNCTION - ONE WHICH IS BEING PERFORMED FOR SUBORDINATE UNITS

Figure 6 Beneficiaries of Functions

The result of beneficiary stratification when applied to a mechanized infantry battalion is presented in Figure 7.

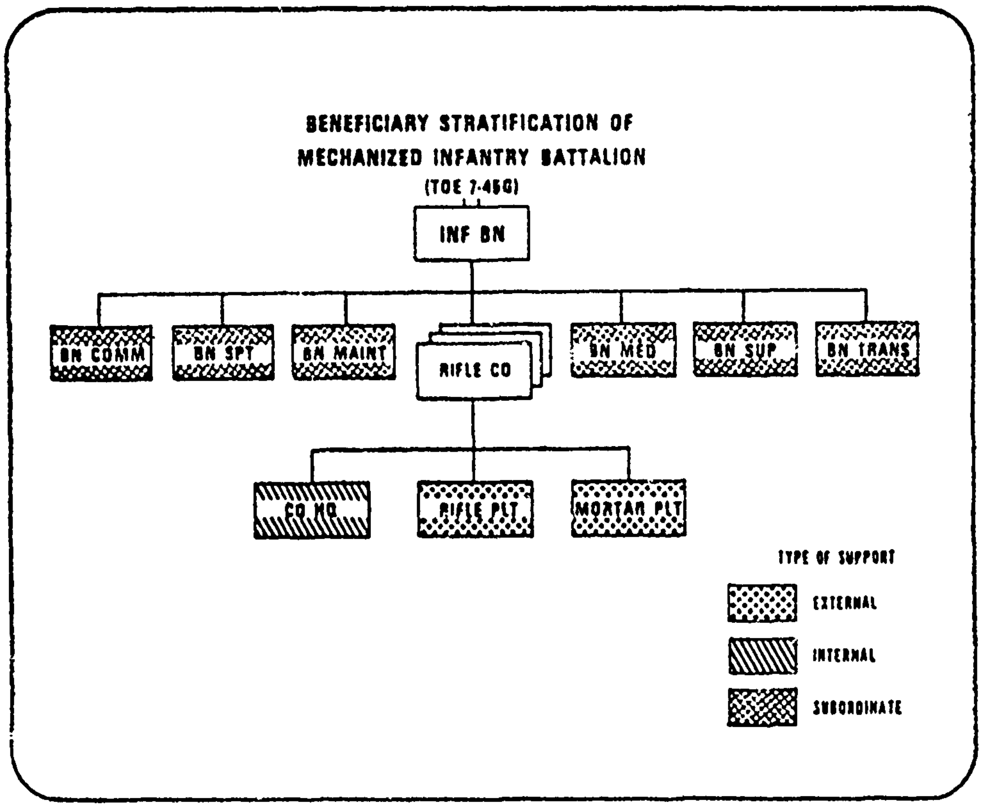


Figure 7 Mechanized Infantry Battalion

2.2.3. The FSS offers many other types of comparison. Figure 8 through 11 illustrate how FSS can be utilized to compare various division forces. These figures show personnel, summary supply, and two categories of the summary supply function, namely ammunition and petroleum, oil, and lubricants (POL).

3. RESOURCE/COST ANALYSIS PROGRAMS

3.1. The Resource/Cost Analysis Programs, (RECAP) were developed to supplement the Comptroller of the Army's Force Cost Information System (FCIS). The FCIS output, examples in Figures 12 and 13, gives total cost for a unit by funding appropriation and theatre but gives little visibility into the force itself.

DIVISION COMPARISON

(MANPOWER IN PERCENT)

FUNCTION	TYPE DIVISION AIRBORNE	INFANTRY	MECHANIZED	ARMORED	AIRMOBILE
I DOE	56.7	51.7	47.9	46.9	39.9
II INTEL	2.0	2.0	2.1	2.1	1.6
III MVT	3.0	3.4	3.4	3.6	7.5
IV COMM	7.2	7.2	6.9	6.9	6.5
V SUPPLY	4.5	4.8	4.7	4.7	6.9
VI MAINT	9.0	12.3	15.1	15.9	19.9
VII ENGR	2.7	3.4	4.6	4.5	2.8
VIII MED	4.3	4.3	4.3	4.3	3.5
IX SVC	10.5	10.7	10.8	10.9	11.2
X MISC	0.1	0.2	0.2	0.2	0.2

Figure 8 Division Comparison

DIVISION COMPARISON

SUPPLY SUMMARY FUNCTION

	TYPE DIVISION ABN	INF	MECH	ARMD	AMBL
NUMBER OF PERSONNEL	682	799	761	782	1,134
RESUPPLY REQUIREMENTS (SHORT TONS/DAY)	872	1,407	2,254	2,557	1,209
WORKLOAD (SHORT TONS/MAN/DAY)	1.3	1.8	3.0	3.3	1.1

Figure 9 Supply Summary

DIVISION COMPARISON

(AMMUNITION SUPPLY)

	TYPE DIVISION				
	ABN	INF	MECH	ARMD	AMBL
NUMBER OF PERSONNEL	25	12	11	11	75
RESUPPLY REQUIREMENTS (SHORT TONS/DAY).	518	901	1,282	1,520	422
WORKLOAD (SHORT TONS/MAN/DAY)	21	75	117	138	6

Figure 10 Ammunition Supply

DIVISION COMPARISON

(PETROLEUM, OIL, & LUBRICANTS (POL) SUPPLY)

	TYPE DIVISION				
	ABN	INF	MECH	ARMD	AMBL
NUMBER OF PERSONNEL	19	33	22	23	107
RESUPPLY REQUIREMENTS (SHORT TONS/DAY)	146	272	692	758	535
WORKLOAD (SHORT TONS/MAN/DAY)	8	8	31	33	5

Figure 11 PSL Supply

VARIABLE COST

	CONUS		EUROPE		PACIFIC	
	NON RECUR	ANNUAL RECUR	NON RECUR	ANNUAL RECUR	NON RECUR	ANNUAL RECUR
Investment Cost						
TOTAL	18249052	1369204	18351206	1392360	18351206	1392360
Direct Cost	18064519	1356718	18166673	1359874	18166673	1359874
Aircraft Proc	0	0	0	0	0	0
Major Equipment	0	0	0	0	0	0
OP Readiness Float	0	0	0	0	0	0
Repair Cycle Float	0	0	0	0	0	0
Rep Paris & Sec Items	0	0	0	0	0	0
Missile Proc	0	0	0	0	0	0
Major Equipment	0	0	0	0	0	0
OP Readiness Float	0	0	0	0	0	0
Repair Cycle Float	0	0	0	0	0	0
Rep Paris & Sec Items	0	0	0	0	0	0
Missiles (ASP)	0	0	0	0	0	0
Proc WPN & TRKD Veh	17493427	5235719	17584254	526344	17584254	526344
Major Equipment	14600098	16338	14600898	16338	14600898	16338
OP Readiness Float	1174066	0	1264893	0	1264893	0
Repair Cycle Float	1134428	0	1134428	0	1134428	0
Rep Paris & Sec Items	584035	507281	584035	510006	584035	510006
Proc of Ammo	0	807090	0	807090	0	807090
Other Proc	571092	26009	582419	26440	582419	26440
Major Equipment	496308	6194	496308	6194	496308	6194
OP Readiness Float	10956	0	22283	0	22283	0
Repair Cycle Float	14198	0	114198	0	14198	0
Rep Paris & Sec Items	49630	19815	49630	20246	49630	20246
Indirect Cost (MOS TNG)	184533	32486	184533	32486	184533	32486
Aircraft Proc	0	0	0	0	0	0
Missile Proc	0	0	0	0	0	0
Proc WPN & TRKD CBT Veh	0	0	0	0	0	0
Proc of Ammunition	184533	32486	184533	32486	184533	32486
Other Proc	0	0	0	0	0	0

Figure 12 Tank Company Investment Cost

VARIABLE COST

	CONUS		EUROPE		PACIFIC	
	NON RECUR	ANNUAL RECUR	NON RECUR	ANNUAL RECUR	NON RECUR	ANNUAL RECUR
Operations Cost						
TOTAL	1901436	2904752	2289351	3444236	2342061	3901004
OMA	714683	754985	1076417	1073880	1029707	1420877
MPA	1186753	2149767	1212934	2370356	1312354	2480127
Direct Cost	557083	2122552	584064	2377592	683484	2633220
OMA Prog 2	360860	223064	360860	257515	360660	403372
PLL Repair Parts	80476	0	00476	0	80476	0
Minor Equipment	176610	0	176610	0	176610	0
ORG Clothing & Equip	103774	0	103774	0	103774	0
Unit OP (Except ACFT)	0	223064	0	257515	0	403372
Aircraft Operations	0	0	0	0	0	0
Other (Static)	0	0	0	0	0	0
MPA	197023	1899488	223204	2120077	322024	2229848
PCS	197023	89367	223204	234230	322024	311745
Other	0	1810121	0	1885847	0	1918103
Indirect Cost	1343553	782200	1705287	1066644	1658577	1267784
OMA (Support & MOS TNG)	353823	531921	715557	816365	668847	1017505
Prog 2 Base OP	0	245284	0	327876	0	458084
Prog 7 (S)	138529	84728	500263	283198	463563	352618
Prog 7 (M)	0	80832	0	80832	0	80832
Prog 8 (M)	2013	50552	2013	57138	2013	50552
Prog 8 (T)	205405	42757	205405	42757	205405	42757
Prog 8 (O)	7876	16554	7876	13350	7876	20648
Prog 9	0	11214	0	11214	0	11214
MPA (MOS TNG)	989730	250279	989730	250279	989730	250279
Investment & Operations						
TOTAL	11340804	3519690	11770464	4060734	11823174	4517502
Direct Cost	9754063	2704158	9821989	2960758	9921409	3216386
Investment	9196180	581606	9237925	583166	90237425	583166
Operation	557883	2122552	584064	2377592	683484	2633220
Indirect Cost	1586741	815532	1948475	1099976	19017654	1301116
Investment	243188	33332	243188	33332	243188	33332
Operation	1343553	782200	1705287	1066644	1658577	1267784

Figure 13 Tank Company Operations Cost

RECAP gives the analyst the ability to look beyond funding appropriations and evaluate the specific cost of personnel and/or equipment by MOS and line item number (LIN) down to the paragraph level of a force.

3.2. The RECAP process consists of two programs, the cost program (CSTSRC) and the print program (PRTSRC).

3.2.1. The cost procedure, outlined in Figure 14, is a program which develops a unique LIN and MOS list from the force structure specified; retrieves the cost for each LIN and MOS and builds direct access files which contain the unique LINs and MOSs with their associated nonrecurring and recurring cost.

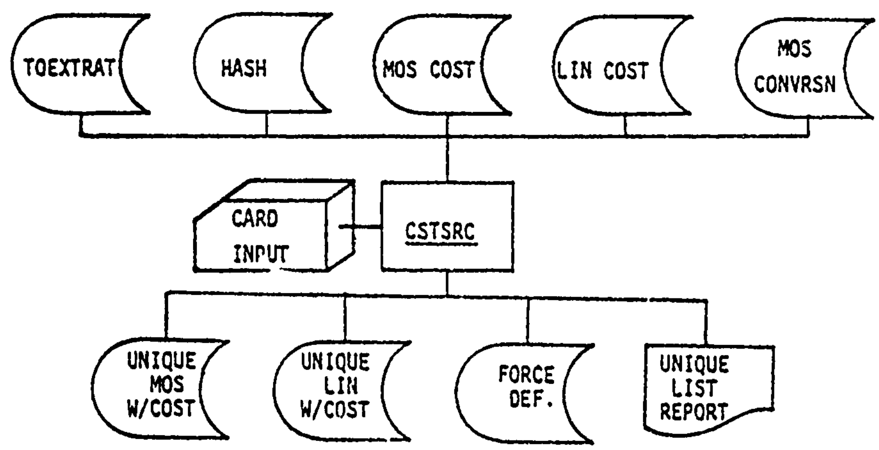


Figure 14 RECAP Cost Program

3.2.2. The print program, outlined in Figure 15, is a set of programs which print out SRCs (forces) at the MOS and LIN level with their respective nonrecurring and recurring costs.

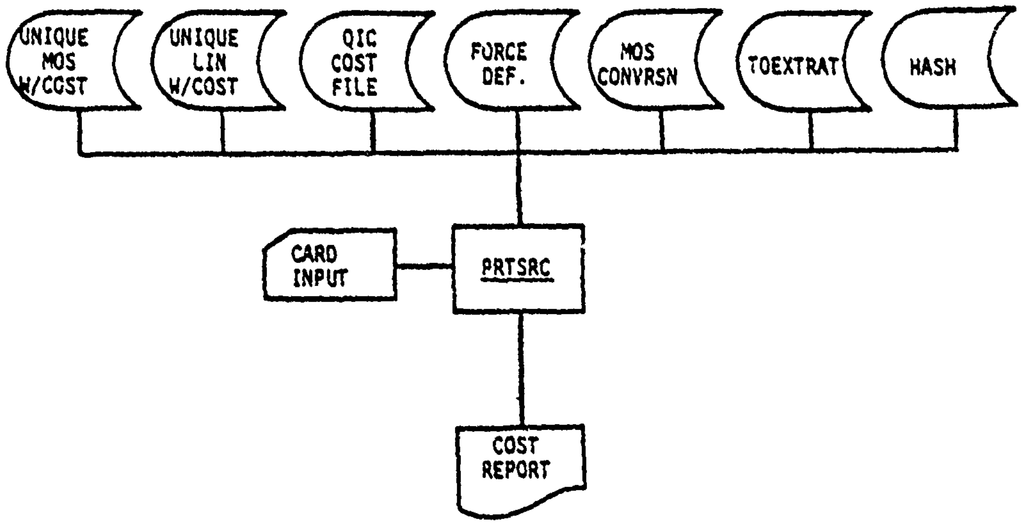


Figure 15 RECAP Print Program

3.3. EXAMPLES OF RECAP OUTPUT

3.3.1. In an Infantry System Special Study Group report, RECAP was used to provide two views of the cost for five alternative divisions. These costs are total 20-year costs. Full division costs imply total TOE costs, i.e., for all units in the force costs are computed for all personnel and equipment. Variable division costs imply costing only those units which directly contribute to the destruction of the enemy. Figure 16 summarizes the SSG division costs.

Division Alternative	Full Costing	Variable Costing
1	\$14.5	\$2.4
2	\$14.3	\$2.3
3	\$14.7	\$2.5
4	\$14.4	\$2.4
5	\$14.1	\$2.4

Figure 16 SSG 20-Yr Division Cost FY 78 Billion \$

3.3.2. As is apparent from Figure 16, force cost discrimination is not apparent at the division level. More often, RECAP is used to evaluate forces of battalion or company size. Figure 17 presents the force cost summary obtained from RECAP for a tank company team defensive scenario.

3.3.3. Costs input in both Figures 16 and 17 are built from the platoon level up. Herein lies one of the greatest assets of RECAP. At the platoon level, RECAP allows for the automated substitution, addition, and deletion of both personnel and equipment.

Company Team Defense	Tk1		Tk2		Tk3		Tk4	
Maneuver Force	NR	AR	NR	AR	NR	AR	NR	AR
1 Tk Co HQ	1,810	481	2,142	498	2,788	524	2,869	567
2 Tk Plt	7,483	1,489	9,143	1,577	12,374	1700	12,769	1,870
2 Mech Plt	4,359	1,492	-		-		-	
2 At Sec	1,779	396	-		-		-	
SUBTOTAL	15,431	3,858	17,423	3,963	21,300	4,112	21,776	4316
20-Yr TOTAL	$92,591		$96,683		$103,540		$108,076	
Supporting Forces								
2 AD Sec	8,887	1,215						
2 STINGER Tm	147	96						
5/7 Atk Helo Plt (-)	31,456	2,207						
1/9 Scout Helo Plt (-)	14,248	929						
FIST	317	177						
SUBTOTAL	55,055	4,624						
20-Yr TOTAL	$147,535							
Support Slice for Maneuver Force								
2/3 Tk Co Maint Sec	1791	664	1876	663	2036	675	2,056	683
1/6 Tk BN HHC	779	187	-		-		-	
2/3 Mech Co Maint Sec	732	163	-		-		-	
1/6 Mech BN HHC	1315	607	-		-		-	
SUBTOTAL	4617	1,621	4702	1625	4862	1632	4882	1640
20-Yr TOTAL	37,037		37202		$37,502		$37,682	

Figure 17 FORCE COST SUMMARY
FY 79 K

4. SUMMARY

4.1. In summary, Force Stratification has utility as a planning tool for force structuring and unit design. It can be used to examine the allocation of resources for unit, battalion, division, corp, theater, or total Army forces. It is often used to answer questions on combat to support ratio; and resources allocation between functional areas. However, it is most often used to locate particular personnel (by functional area or MOS or ASI) or equipment (by LIN).

4.2. The future of Force Stratification depends on the users' desires. Some users have voiced a desire to make the system accept MTOE units. These are actual units as they have been modified by field commanders to suit their particular missions. This change requires some modification to the system but can be accomplished if this need is genuine. The system has already been modified to accept conceptual units composed of either present personnel and equipment or conceptual personnel and/or developmental equipment. This modification is very useful in evaluating futuristic forces.

4.3. RECAP, like FSS, is a resource model designed to assist the analyst in performing a force analysis. As mentioned earlier, it was designed to compliment the output of FCIS by providing the analyst with visibility into the actual force structure. These models will not perform the analysis for the analyst but enables the analyst to examine large volumns of data in a timely manner without becoming bogged down in the detail of the force structure.

4.4. The next ten years is going to be a period seeing countless new systems entering the Army inventory. The resource implications of each of those new systems will have been evaluated with respect to the present force structure, but what about the force structure when all of these new sophisticated systems are in the field? FSS and RECAP are tools which can assist the analyst in objectively evaluating these many resource interactions since the models have the ability to evaluate very large forces in a timely manner.

POLITICO-MILITARY SIMULATION METHODOLOGY

JAMES B. MOTLEY
LIEUTENTANT COLONEL, U.S. ARMY

Politico-Military Division
Studies, Analysis, and Gaming Agency
Organization of the Joint Chiefs of Staff
Washington, D.C. 20301

ABSTRACT. This paper will discuss the mission of the Politico-Military Division, Studies, Analysis, and Gaming Agency (SAGA), Organization of the Joint Chiefs of Staff. The purpose and methodology of how SAGA utilizes politico-military (PM) simulations as an interagency tool within the executive community to openly discuss international security issues, examine controversial questions, and propose innovative solutions to policy problems will be explained.

PM simulations are defined as simulation of situations involving the interaction of political, military, sociological, psychological, economic, scientific, and other appropriate factors.

An example of the flow of events for a three move simulation will be presented.

PRESENTATION ON
POLITICO-MILITARY SIMULATION METHODOLOGY
PACIFIC CONFERENCE ON OPERATIONS RESEARCH
(23-25 APRIL 1979)
SEOUL, KOREA

Ladies and Gentlemen:

On behalf of the Chairman of the Joint Chiefs of Staff and the Chief of the Studies, Analysis, and Gaming Agency I would like to express my appreciation for the opportunity to participate in this conference.

The purpose of my presentation is to discuss how the Studies, Analysis, and Gaming Agency, commonly referred to as SAGA, a separate agency within the Organization of the Joint Chiefs of Staff (OJCS), utilizes politico-military (PM) simulations. My comments will be unclassified. Time permitting there will be a question and answer period upon completion of my remarks.

The Politico-Military Division, Studies, Analysis, and Gaming Agency, OJCS has a unique mission in that it is chartered to conduct politico-military simulations on an interagency basis, thereby serving the entire executive branch of the U.S. Government. The simulations are approved and sponsored by the Chairman of the Joint Chiefs of Staff and address problems with broad national security implications. These simulations are the only vehicle of their kind conducted in the U.S. Government on a continuing basis.

This is the formal definition of politico-military simulations. "Simulation of situations involving the interaction of political, military, sociological, psychological, economic, scientific, and other appropriate factors." The simulations are intended to: identify major international issues and problems; provide a forum for examination of controversial policies; encourage the exchange of ideas; and surface new or innovative approaches to the problems and issues thus identified. While care must be exercised in applying the results of simulations to

real-world circumstances, politico-military simulations do provide participants an opportunity to examine and debate, in an unencumbered environment, major national security problems.

Because of the importance attached to the teaching aspects of the methodology, politico-military simulations are conducted on a regular basis in the Pentagon and also at several overseas locations, thus allowing wide exposure to this type of crisis management exercise. I might add that all of the senior service colleges conduct politico-military simulations as crisis management training vehicles.

Politico-military simulations tend to serve these purposes. By identify we mean to bring a possible trouble spot to the attention of the players, not necessarily to predict a future crisis.

After having identified a future trouble spot, it is hoped that new ideas and new approaches to crisis management are surfaced. The simulations then serve as a useful means to test these new ideas and approaches to crisis management simulations. The simulations also serve as a useful means to test new ideas and approaches without the prospect of paying real-life consequences for mistakes or misjudgments.

In trying to establish useful solutions to a crisis, participants address significant issues, problems and questions of the region. Inasmuch as scenario time is set some two-three years in the future, and SAGA's adherence to a strict policy of nonattribution, participants are free to suggest new alternatives that may well be counter to present policies. Each simulation is a forum for the free exchange of ideas. When the simulation is completed, the collective scenario that has evolved often assists in evaluating ongoing planning efforts.

In addition politico-military simulations help facilitate communications as participants share a common crisis management experience. Being exposed to the decision-making process and the interrelationships of diplomatic and

military actions, players cannot go through a simulation without acquiring a large body of information that will assist them in their day-to-day work. More importantly, the simulation experience may assist them in handling crises they may face in the future and in understanding how other agencies manage crises.

At this time I will present a short movie which provides an overview of the PM technique, as employed by SAGA. I will then discuss in greater detail an explanation of the mechanics of the PM technique -- its requirement, costs, variations and productivity.

I trust that the film has provided some insight into the requirements and techniques of PM simulations.

As noted in the film, SAGA's PM simulations are completely manual and do not involve the use of computers. Generally, policy decisions -- especially in times of crisis -- result from collective judgment rather than quantifiable knowledge. PM simulations are designed to emulate this decision-making process by assuring freedom and spontaneity in team discussions and decisions.

An absolute requirement for a successful PM simulation is a challenging scenario. As shown on this slide, the staff planning for the exercise has many directions it can move in selecting the focus of the scenario. The simulation objectives must influence this selection. For example, a simulation designed to test existing policy will differ from one intended to evolve new policy. The scenario then, is the key element in the simulation since it serves as the basis for player team discussions and subsequent decisions.

As indicated in the film, PM simulations are built around player teams. Each team is composed of approximately ten participants. Participants are assigned to the player and control teams so as to achieve a balance in regional and or technical expertise. For example, a team might well include representatives from the military services, intelligence agencies, the President's staff, the economic

community and the diplomatic corps. Thus, the team is capable of developing options from a broad base of knowledge and experience.

Players are invited from all segments of the executive branch and from the academic community. In this manner, a politico-military interface occurs -- diplomats learn from the generals, economists learn from the politicans and so forth.

Both senior and intermediate-level simulations are conducted throughout the year. Although the grades of participants range from colonel to cabinet member, this is not the critical factor. What is important is the depth of expertise and knowledge on the selected topic that the participant can bring to the team discussions.

The film depicted an adversary relationship between the red and blue teams. In this structure, the members of the blue team represented the National Command Authority of the United States, the members of the red team the Politburo of the USSR, and the members of the control team all other nations and international organizations. It is important to note that in our games individual participants do not role-play their real life positions. For example, in the case of the blue team, they represent the corporate U.S. National Command Authority and their actions are presidential decisions.

This structure is a variation of the red-blue game. A third player team, green, is added and represents either a nation friendly to blue or an international organization such as NATO. Although any number of player teams can be added to the game, the greater the number, the more difficult the simulation.

A third variation eliminates the red team. In this parallel relationship, both blue and green represent the same nation, or the same international organization, and control represents everyone else. In essence, two simulations are running simultaneously, both starting with the same basic scenario but likely to diverge as the game progresses and the blue-green actions differ. In this

structure there is no interaction between the blue and green teams. In terms of net result, we believe this latter structure is the most productive and cost effective. Two separate approaches to a common crisis, are obtained which can be compared for their new ideas and effectiveness, all for the efforts of one staff, one scenario and one control team.

The exercise commences with the player teams reading a background scenario and responding to the scenario crisis. No one is allowed in the team room during game play except team members and one technical representative from the PM Division of SAGA who provides assistance to the team. The team members require no ancillary support. The teams develop a message as outlined on this slide, in response to the crisis.

The team assesses the situation, determines its objectives and selects the specific political, military, economic and other actions appropriate to the strategy. A contingency paragraph concludes the team messages; it provides control with some "expected actions" by the player team. These are useful if control projects time more than a few hours or days.

It should be noted that a politico-military simulation is not a war game; the emphasis is not on the movement of military forces or command and control. In PM simulations the emphasis is on the whole picture rather than purely military measures. This approach also differs from Command Post Exercises (CPX's) in a number of ways. In a PM simulation, the interaction of diplomatic, economic and military factors is addressed and vividly impressed in the minds of the participants. It is felt that the free-form structure and the fact that there are no pre-determined actions, as in the case of the CPX, lead to the formulation of new ideas and approaches to problems.

The scenario is written only to the point of the first team meeting. Scenario projections for subsequent team meetings are totally dependent on the actions taken by the player teams as accepted and interpreted by the

control team. Control literally writes the projections, one for each player team.

In the flow chart shown here, moving from left to right, you see that the control team meets after the teams have met and passed their message or decisions to control. Control assesses each team message and prepares a scenario projection outlining a new situation which will develop some hours or days later and which will require the teams to reassess the situation and undertake further actions. This process is repeated for move III. After all moves have been completed, a critique is held which provides participants an opportunity to exchange opinions, comment on scenario events, and discuss related "real life" questions. This critique period is considered the "highlight of the exercise" and a major source of information for all concerned.

In SAGA's simulations, control monitors player team discussions on closed circuit television (TV) to hear the rationale leading to team decisions. We consider the TV monitoring essential for that purpose and also to hear minority or dissenting opinion.

Although experience has indicated that a three move simulation is the most productive, we have tried a number of variations. Our "one-move", "snap-shot" game is basically limited to a scenario situation in which the participants are forced to make a single, major decision. A two move exercise can be productive and is recommended when time is a factor -- it can be completed in two days.

Our three move games usually run four or five days. This is the type game we use most often. The player team participants are involved for a two hour introductory period and three four-hour team meetings, plus the critique which lasts about two hours. The control team has two additional four hour sessions and a one hour session.

Let us now look at how the Politico-Military Division organizes to prepare for and conduct a simulation. Our primary members

of the task force are the project officer, the scenarist and the conference manager. The project officer has overall responsibility for the simulation. He is assisted in this role by the scenarist, who, in addition to writing the scenario, is also the technical representative on the control team. The project officer and conference manager are responsible for all administrative functions, such as site preparation, invitations, supplies, production and transportation. In addition, a staff officer is assigned to each player team and, like the scenarist, provides technical assistance to his team. This player team staff officer's total obligation is about six weeks -- four weeks prior to the simulation for research; one week during the simulation play; and one week for post game documentation.

If you decide to conduct a PM simulation, I would suggest this type of task force organization. Four staff officers can make all game preparations and write the scenario. They will, however, require additional assistance just prior to and during game play consisting normally of six secretarial and six enlisted personnel.

Time required for post-simulation documentation and analysis will depend on the objectives set for the game and will normally require the full time effort of one staff officer for approximately a month.

At this time I will be glad to answer any questions you may have on PM techniques and methodology. I will also be available after this morning or afternoon session to sit down and discuss SAGA's methodology in more detail with you on an individual or group basis.

AN APPLICATION OF STOCHASTIC DYNAMIC PROGRAMMING FOR DETERMINING THE OPTIMAL REPAIR LIMITS FOR A FLEET OF VEHICLES

H. PAUL

Department of Industrial & Systems Engineering
University of Singapore
Kent Ridge, Singapore 5
Republic of Singapore

ABSTRACT. A policy based on certain repair limits corresponding to various ages of vehicles may be used to determine whether a vehicle should be repaired or replaced. Under such a policy, when a vehicle requires repair it is first inspected and the total repair cost is estimated. If the estimated repair cost exceeds the repair limit corresponding to the age of the vehicle, the vehicle is not repaired but replaced. The set of repair limits corresponding to various ages of the vehicles provides an economic replacement policy which ensures that the vehicles are continuously replaced as soon as they become uneconomical to maintain. Though this policy is a useful one, the set of repair limits must be determined systematically to ensure that an optimum policy is followed. This paper establishes a method, based on the application of dynamic programming to Markov processes, of determining an optimum set of repair limits for a fleet of light vehicles. The method is illustrated by an example to show the savings realized if an optimum policy based on the repair limits is followed.

LIST OF SYMBOLS

C	Initial cost of a vehicle
E_i	Expected duration of stay in state i
$f(x_i)$	Probability density function of repair costs
$F(x_i)$	Probability function of repair costs
g	Average annual cost of repair and replacement for the whole fleet of vehicles
k_i	Failure rate of vehicles in state i
K	Mean failure rate of vehicles for all states
N	Maximum age of vehicles
p_{ij}	Probability that a vehicle will go from state i to state j
q_i	Expected cost in the next transition out of state i
s_i	Probability of survival at state i
v_i	Relative value of the system at entering into state i
x_i	Repair limit in state i
λ_i	Parameter of the repair cost probability density function

1. INTRODUCTION

1.1 Background

An organization uses different fleets of vehicles for various purposes. Within each fleet there are vehicles of different ages. When a vehicle is found to be uneconomical to repair it is replaced by a similar type of vehicle.

Currently, the organization uses a replacement policy for vehicles in which for each vehicle fleet a set of repair limits corresponding to the age of the vehicles is determined rather arbitrarily by depreciating the value of each vehicle by a constant factor of 0.85 for each succeeding year. Under such a policy when a vehicle requires repair, it is first inspected and the total repair cost is estimated. If this estimated repair cost exceeds the repair limit corresponding to the age of the vehicle, it is not repaired but replaced. The set of repair limits thus provides an economic replacement policy which ensures that the vehicles are continuously replaced as soon as they become uneconomical to be maintained.

While this is a useful policy, particularly where replacement can be determined solely on economical factors, the set of repair limits for the various ages of a fleet of vehicles is not determined systematically to ensure that an optimum policy is followed.

The organization maintains a chronological record of all repairs done on every vehicle in a maintenance record book. From such records, the relationship between age and cost of maintenance can be investigated for every vehicle. This age related cost data can be used to determine an optimum replacement policy for each fleet of vehicles.

The aim of this paper is to establish a method whereby such data can be used to determine an optimum set of repair limits for a fleet of light vehicles. These repair limits can be used in the replacement procedure for these vehicles.

1.2 Literature Survey

An economic replacement model using repair limit theory was first investigated by Drinkwater and Hastings using simulation and analytical methods [1]. The simulation method was carried out using a computer and the optimum policy was obtained after 4 iterations and each iteration took 15 minutes on the Leo KDF 9 computer. The mathematical method was more efficient taking only 5 minutes on the same computer using the same set of data.

Hastings made further contributions to the development of repair limit replacement problems using dynamic programming [2, 3]. In his studies he had made assumptions that renewal programming can be used to formulate and solve these repair limit replacement problems.

Two main types of problems were analysed. In one the equipment condition was related to age and in the other the equipment condition was related to the number of repairs. In the analysis of these problems, Hastings had assumed an Erlang distribution for the repair costs and a constant failure rate.

Mahon and Bailey proposed an improved replacement policy for army vehicles for the British Army [4]. The model used was essentially that proposed by Drinkwater and Hastings with a little modification [2]. In this model proposed by Mahon and Bailey the cumulative cost of repairing and replacing a vehicle and its successor was used, and this cost was represented by a stochastic process in continuous time. Using REME's FORWARD data base, the authors set out to find a suitable mathematical formula for the repair costs and repair frequency probability distributions. From very large data samples they concluded that best fit for the repair frequency was a family of Newman type A distributions and that for the repair costs it was a family of Weibull distributions.

The optimum repair limit policy was determined using the dynamic programming method developed by Drinkwater and Hastings [2]. Calculation was done on an ICL 1904 computer in about 5 minutes.

2. FORMULATION OF THE MODEL

2.1 The Recurrence Relationship

The problem of establishing repair limits can be analysed using the theory of Markov Process with rewards or costs. In a continuous Markov Process, the recurrence relationship may be stated as[1]

$$v_i = -gE_i + q_i + \sum_{j=1}^{N} p_{ij}v_j \qquad (1)$$

For a particular state i, only two types of transitions

---

1. For details of derivation of the relationship, refer [6].

are possible. The vehicle can either survive (if the repair cost is lower than the repair limit) or be replaced (if the repair cost exceeds the repair limit).

From equation (1)

$$v_i = -gE_i + q_i + (1 - s_i)v_i + s_i v_{i+1} \qquad (2)$$

The variables E_i, q_i and s_i can be expressed as functions of the repair limit x_i. These functions will depend on the probability distribution of the repair costs and repair frequencies.

A boundary condition on the age of the vehicles in terms of maximum age must be stated. If this maximum age is N, then when the vehicle reaches year N, it will be replaced. Hence in the Nth year

$$v_N = C + v_1 \qquad (3)$$

2.2 Assumptions

It is assumed that the failure rate and other properties of the vehicle is related to age. This age related model is valid provided that the vehicle consists of a fairly large number of components and repair work is carried out to overcome a particular malfunction and not to raise the item to a standard of reliability significantly higher than the average of its components [3].

The cost of unavailability and cash discounting rate have been excluded from the formulation. These factors have opposite effects in the determination of average optimum life of vehicles; one lengthens the average vehicle life whereas the other shortens it [4].

The recurrence relationship is applicable only if the system is in a steady state. Steady state is achieved if the replacement policy is followed for some period of time.

2.3 Solution of the Recurrence Relationship

Let the repair cost distribution function is $f(x_i) = \lambda_i e^{-\lambda_i x_i}$. The probability that the repair cost will exceed x_i, is $p_{ij} = e^{-\lambda_i x_i}$

Let $M_i = \int_0^{x_i} x_i\, f(x_i)dx_i$, then

$$M_i = (^1/\lambda_i)(1 - p_{ij}) - x_i p_{ij} \qquad (4)$$

The following relationship can be derived [8]:

$$s_i = e^{-p_{ij}k_i} \qquad (5)$$

$$E_i = (1 - e^{-p_{ij}k_i})/p_{ij}k_i = (1 - s_i)/p_{ij}k_i \qquad (6)$$

$$q_i = (^{M_i}/p_{ij} + C)(1 + s_i) \qquad (7)$$

2.4 Repair Cost Distribution

The fleet of light vehicles considered in this problem consists of 700 units. Data were obtained from the maintenance records of a sample of 100 of these vehicles. The data available in the maintenance records were the spare parts cost and the manhours spent on each repair and the date on which the repair was carried out. To obtain the total cost of a repair a rate of \$10 per manhour was used to compute the labour cost. Since the rates charged by automotive repair firms are normally in the \$10 to \$12 range and some of the repair jobs contracted out were being charged at \$12 per manhour, the rate of \$10 used was considered reasonable. Furthermore, the overhead and indirect labour incurred were generally lower.

A good fit was obtained using an exponential distribution function between repair costs of vehicles of a given age [8]. Similar conclusions were obtained by Hastings [1]. The probability function is given by

$$F_i = 1 - e^{-\lambda_i x_i} \qquad (8)$$

λ_i is a function of the age of the vehicles. The values of λ_i's for the first 15 years of vehicle life were estimated for the sample of 100 vehicles. These values are given in Table 1.

Table 1 Parameters for Repair Costs Distribution Functions for Various Ages of Vehicles

Age (Years)	λ_i	Age (Years)	λ_i
1	0.00780	9	0.00120
2	0.00560	10	0.00114
3	0.00380	11	0.00102
4	0.00250	12	0.00092
5	0.00180	13	0.00083
6	0.00160	14	0.00077
7	0.00150	15	0.00074
8	0.00136		

2.5 Repair Frequency Distribution

The number of repairs done on each vehicle for each year of its life were obtained from the maintenance records of the same sample of 100 vehicles. It was noted that there is very little variation in the repair frequency. The mean failure rates were computed for each of the 15 age groups. These mean failure rates for the various age groups are shown in Table 2. It was noted that the mean failure rates at age 6 years and above vary very slightly from year to year, and in the last five years the variation is insignificant. This small variation was probably due to the fact that every vehicle is given a thorough inspection bimonthly and nearly all repairs were done after this inspection.

Table 2 Mean Failure Rates for Various Ages of Vehicles

Age (Years)	k_i	Age (Years)	k_i
1	3.21	9	4.63
2	3.62	10	4.74
3	3.74	11	4.80
4	3.97	12	4.81
5	4.15	13	4.83
6	4.37	14	4.84
7	4.42	15	4.86
8	4.51		

To greatly simplify the mathematical expressions for the variables used in the recurrence relation given in equation (2), the failure rate was taken to be a constant and equal to the mean failure rate K. This assumption of a constant failure rate was acceptable as the variation of the repair frequency within each age group is small and the mean fail ure rates for various age groups increase very insignificantly with time for vehicles of age 10 years to age 15 years.

There may be some concern over the values of the mean failure rates for the first six years as the mean failure rates vary more than those for the latter years. However, the results of the repair limits obtained for the first eight years were of no practical significance as the probability of replacing a vehicle at this age was almost equal to zero. The optimum value of repair limit for each year is independent of those before and after it. Hence the results of the first eight years will have no effect on the policy for the remaining years.

3. APPLICATION OF THE MODEL

3.1 Maximum Age of Vehicles and Accuracy of Repair Costs

For the model developed in the previous section, the maximum age for the fleet of vehicles must be stated. The maximum age for the fleet under consideration was taken to be 16 years. This is because it was found from statistics that the type of vehicles in the fleet became uneconomical to maintain at about 14 years of age and the vehicles are normally replaced within the next one or two years upon reaching this age.

To limit the large number of possible repair costs, the repair costs were considered in steps of S$50. This is acceptable because in practice it is impossible.to estimate a repair cost exceeding S$500 to within S$50, and all repair limits exceed S$900 [8].

3.2 Value Iteration

Using N = 16, the recurrence relation from equations (2) and (3) is

$$v_i = -gE_i + q_i + (1 - s_i)v_i + s_i v_{i+1} \qquad i = 1, 2, \ldots 15 \qquad (9)$$

$$v_N = C + v_1 \qquad (10)$$

These equations (sixteen in total) have a total of 17 un-

knowns, viz., g and v_i's $i = 1, 2, \ldots 16$.

At state age 15 years

$$v_{15} = -gE_{15} + q_{15} + v_1 + s_{15}C \qquad (11)$$

Hence equations (9) and (10) are now reduced to 15 equations with 16 unknowns, namely, g and v_i where i = 1 to 15.

A set of values of repair limits, x_i's is first estimated. These values are used to determine the values of v_i's by solving the simultaneous equations (9) and (11). This is the value iteration procedure.

3.3 Policy Iteration

Using the new set of v_i's the optimum policy (given by a new set of repair limits, x_i's) is determined by minimising each v_i for various values of the corresponding repair limit x_i. Minimising the set of v_i's ensures that the cost of maintaining the fleet is minimum. This is the policy iteration procedure.

3.4 Computational Procedure

Using the optimum policy (new set of x_i's) determined in the policy iteration procedure, the value iteration procedure is repeated. The policy-value iteration procedures are continued alternatively until there is no improvement in the policy between successive iterations.

To begin the computations, the values of the 16 unknowns of equations (9) and (11) were first estimated. For each state i, the values of E_i, q_i, and s_i were computed for various values of x_i. The value of x_i that minimises the right hand sides of equations (9) and (11) is the repair limit for that state.

The procedure was repeated for all i = 1 to 15 and the set of x_i's so obtained give the optimum policy.

With the optimum policy derived, equations (9) and (11) were solved for the values of v_i's and g. As discussed above, the recurrence relation has 15 equations with 16 unknowns. To solve this set of simultaneous equations, v_1 was set equal to zero. In the determination of the optimum repair limits, the absolute values of v_i's are not of interest.

Hence in this analysis, only the relative values of v_i's were determined.

The relative values of v_i's for i = 2 to 15 and the absolute value of g were hence determined using the values of E_i, q_i and s_i corresponding to the optimum policy.

The above procedures for policy and value iterations were repeated alternatively until no change in the optimum policy was noted. The optimum policy was arrived at after four iterations.

It was found that the solution of the set of 15 simultaneous equations consisting of 15 unknowns can be handled by the elimination method as each equation has at most three unknowns, and by systematic elimination, the equations can be solved fairly easily.

All computations were manually done using desk calculators and it was found that the calculations were manageable without the use of a computer.

3.5 Results

The results of the policy and value iteration procedures are shown in Table 3. It was found that the optimum repair limits for the first five years could not be determined by desk calculators as these limits are greater than S$6000 and the probability that any repair cost will exceed these repair limits were smaller than 0.00001. Also, the repair limits for vehicles of age below eight years cannot be accurately determined, and were not reliable. This was because of the inaccuracy involved in the computation of the exponential function $e^{-p_{ij}K}$ when the product $p_{ij}K$ is small, and also it was necessary to take the difference of two numbers which were both nearly equal. However, the accuracies improved rapidly for repair limits above age nine years.

The above procedure for determination of the optimum repair limits, and the use of manual calculation were acceptable because only the repair limits of vehicles beyond age ten years were important as no vehicle was ever replaced before reaching age ten years, except in cases of accident. The replacement policy developed here does not apply to vehicles involved in accidents causing considerable damage to the vehicle.

Table 3 Policy-value Iterations

Iteration 1			Iteration 2			Iteration 3			Iteration 4		
$g = 5000.0$			$g = 4050.4$			$g = 4031.2$			$g = 4028.9$		
Age i	Values of v_i	Repair Limit x_i	Age i	Values of v_i	Repair Limit x_i	Age i	Values of v_i	Repair Limit x_i	Age i	Values of v_i	Repair Limit x_i
1	0	-	1	0	-	1	0	-	1	0	-
2	3000	-	2	3650	-	2	3104	-	2	3626	-
3	6000	-	3	7080	-	3	6428	-	3	7033	-
4	9000	-	4	10166	-	4	10094	-	4	10095	-
5	11000	-	5	12687	-	5	12591	-	5	12595	-
6	13000	5150	6	14446	5150	6	14327	5150	6	14328	5150
7	14500	4800	7	15766	3650	7	15638	3850	7	15640	3850
8	16000	3850	8	16871	2800	8	16736	2900	8	16737	2900
9	17000	3150	9	17623	2250	9	17499	2300	9	17500	2300
10	17500	2600	10	18000	1900	10	17921	1900	10	17921	1900
11	18000	2000	11	18313	1650	11	18291	1650	11	18291	1650
12	18500	1650	12	18454	1500	12	18462	1500	12	18461	1500
13	19000	1450	13	18597	1400	13	18606	1350	13	18604	1350
14	19300	1400	14	18702	1250	14	18702	1250	14	18703	1250
15	19500	1200	15	18894	950	15	18866	950	15	18867	950

Table 4 gives the values of E_i, q_i and s_i corresponding to the optimum set of repair limits obtained for the fleet of vehicles for ages six years and above. From the optimum policy determined, it can be seen that at age 15 years the repair limit is S$950, the probability of survival is only 0.09 and the expected duration of stay in that state is only 0.378. These figures confirm the choice of a maximum age of 16 years as reasonable.

Table 4 Values of s_i, E_i, q_i, and x_i for the Optimum Policy

Age Years	Optimum Repair Limit x_i, S$	Probability of Survival s_i	Expected Duration of Stay E_i	Cost Associated with State i q_i
6	5150	0.9988	1.0000	2744
7	3850	0.9864	0.9932	3138
8	2900	0.9140	0.9563	4602
9	2300	0.7460	0.8668	7629
10	1900	0.5868	0.7751	10315
11	1650	0.3978	0.6533	13584
12	1500	0.2782	0.5800	15253
13	1350	0.2070	0.5035	16764
14	1250	0.1575	0.4558	17568
15	950	0.0902	0.3781	18586

The value of average annual cost of repair and replacement per vehicle for the fleet is S$4029 under the optimum policy. Inaccuracies in the computation of the variables for the first eight years of the life of the vehicles do not affect this cost significantly.

4. CONCLUSIONS

A model using policy-value iteration method of computation derived from dynamic and Markov renewal programming is obtained. Data gathered from maintenance records on a sam-

ple of 100 vehicles in a fleet of 700 light vehicles suggest that an exponential distribution can be used to represent the repair costs distribution and the yearly failure rate can be reasonably considered to be constant.

If one is not concerned about the inaccuracies of the repair limits obtained for the first half of the vehicle life cycle, the model can be used to determine an optimum repair limit policy quickly and efficiently. These inaccuracies in practice are acceptable as replacements occur only during the later part of a vehicle life cycle.

The repair limits for the present replacement policy are obtained by multiplying the previous your repair limit by a constant factor of 0.85. The average annual cost of repair and replacement for each vehicle under the present policy was found to be S$4112. Thus a saving of over S$58,000 per year is realised if the optimum policy determined by the model proposed in this paper is applied to the whole fleet of 700 vehicles.

The method described in this paper is applicable to all replacement problems involving equipment whose cost of repair and replacement is related to age.

5. REFERENCES

[1] Hastings, N.A.J., SOME NOTES ON DYNAMIC PROGRAMMING AND REPLACEMENT, Operational Research Quarterly, Vol. 19, No. 4, pp 453-464, December 1968

[2] Drinkwater, R.W. and Hastings, N.A.J., AN ECONOMIC REPLACEMENT MODEL, Operational Research Quarterly, Vol. 18, No. 2, pp 121-138, June 1967

[3] Hastings, N.A.J., THE REPAIR LIMIT REPLACEMENT METHOD, Operational Research Quarterly, Vol. 20, No. 3, pp 337-349, September 1969

[4] Mahon, B.H. and Bailey, R.J.M., A PROPOSED IMPROVED REPLACEMENT POLICY FOR ARMY VEHICLES, Operational Research Quarterly, Vol. 26, No. 3, i, pp 477-494, September 1975

[5] Howard, R.A., DYNAMIC PROGRAMMING AND MARKOV PROCESSES, John Wiley, New York, 1960

[6] Jewell, W.S., MARKOV RENEWAL PROGRAMMING, Operations Research, Vol. 11, pp 938-971, November-December 1963

[7] Cox, D.R., RENEWAL THEORY, Methuen, London, 1962

[8] Chia, P.C., AN OPTIMAL REPAIR LIMIT REPLACEMENT POLICY FOR LIGHT VEHICLES, Unpublished M.Sc. Dissertation, University of Singapore, Singapore, 1978

AN ALGORITHM FOR INEFFICIENT REPAIR POLICIES

TAPAS K. SARKAR

Department of Econometrics and Operations Research,
Monash University, Wellington Road,
Clayton, Vic 3168, Australia.

ABSTRACT. In the age replacement model of Barlow and Proschan, a stochastically failing equipment, with increasing failure rate, is replaced (or repaired with 100% efficiency) at age T or earlier failure. After the action the age of the equipment becomes zero and the failure rate of the equipment remains the same as the original equipment at identical ages.

In this paper inefficient repair has been considered. After such an action the age of the equipment becomes zero, but its failure rate at any age x will be higher than that of the original equipment at the same age. It has been assumed that the equipment will be repaired inefficiently $n-1$ times and then will be replaced by an equipment identical with the starting one. At every decision point various repair options will be available. Using renewal theory, long-term average cost per unit time has been derived and an algorithm has been developed to determine the ages and types of inefficient repair to minimize this cost. In age determination the risk of suboptimality due to the discrete nature of decision points in dynamic programming has been avoided.

Subsequently, sufficient conditions have been developed to find a policy which is optimum over n. The algorithm has also been extended to cover the case where the failure properties of the starting equipment is different from those of the replaced equipments, the latter being all identical.

1. INTRODUCTION

In one of the early models of replacement, Barlow and Proschan [1] have introduced the concept of "age replacement". Under this model a stochastically failing equipment (called original equipment) is replaced at age T or earlier failure. The age T is determined by minimizing the long-term average cost per unit time and the latter is determined by using renewal theory. Instead of replacement, we can consider 100% efficient repair. In either case, after the action the equipment becomes "as good as new", so that the age of the equipment becomes zero and the failure rate of the equipment at any age x remains the same as that of the original equipment at age x .

Sarkar [10, 11, 12] has relaxed the assumption of 100% efficient repair. Under this relaxation, when an action is taken at age T or earlier failure, the age of the equipment becomes zero but the failure rate of the equipment at any age x after replacement may be higher than that of the original equipment at the same age.

Such inefficient repair options can be considered by using the technique of dynamic programming (Barlow and Proschan [2], Hastings [5], Howard [6], Jardine [7, 8], Tahara and Nishida [15], White [16]). However, in dynamic programming, the decisions are taken at discrete points of time and unless the decision intervals are small, suboptimal solutions may be obtained. In this paper we have considered the options of inefficient repair by using renewal theory where decision intervals themselves are decision variables. As a result suboptimality is avoided.

As an illustration let us consider the following example:

Example 1.1: The age at failure of an equipment follows a gamma distribution given by the density function.

$$f(x) = \lambda^2 x e^{-\lambda x}, \ \lambda>0 \ , \ x>0 \ . \tag{1.1}$$

For the original equipment, $\lambda = 1$ and the expected age at failure is 2 . If it is replaced by a similar equipment at age T or earlier failure, the costs are,

(i) $C_{MF}(1,1) = 1100$, if action is on failure,

(ii) $C_M(1,1) = 100$, if action is on survival.

The optimal value of T is 0.6311 and the minimum long-term average cost per unit time is 386.92.

Let us now introduce one more option. The equipment can be repaired inefficiently at an age T or earlier failure such that after repair the age of the equipment becomes zero and the density function of the age at failure becomes that of (1.1) with $\lambda = 1.2$. The expected age at failure is then 1.67 . We can easily verify that the failure rate for $\lambda = 1.2$ is higher than the failure rate for $\lambda = 1$ at corresponding ages.

This inefficient repair action can be taken only once and for a repaired equipment the only option available is replacement by an equipment identical to the original one.

The costs associated with these options are,

(i) $C_{MF}(1,2) = 670$ at inefficient repair on failure,

(ii) $C_M(1,2) = 70$ at inefficient repair on survival,

(iii) $C_{MF}(2,1) = 1700$ at replacement of a repaired equipment on failure,

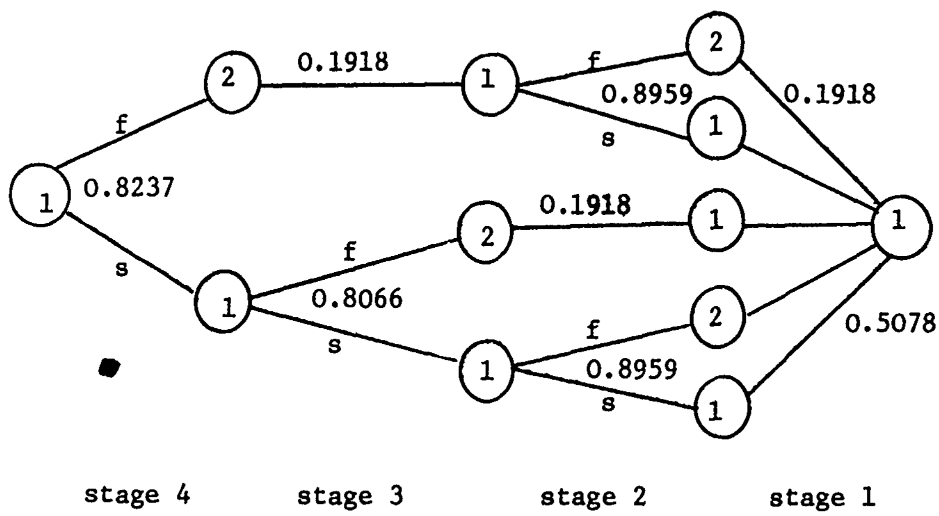
(iv) $C_M(2,1) = 200$ at replacement of a repaired equipment on survival.

Let us use dynamic programming to include the additional option. For decision points at intervals of $t=1$, the optimal policy is to replace the original equipment if it survives to age 1 and to repair the equipment inefficiently if it fails before age 1. For a repaired equipment the optimal action is to replace at age 1 or earlier failure. The long-term average cost per unit time is 388.77.

For decision points at intervals of $t=0.5$, the optimal policy for the original equipment remains the same as for $t=1$. However, for a repaired equipment it is now optimal to replace at age 0.5 or earlier failure. The long-term average cost per unit time is 344.00, 11% cheaper than the age replacement policy.

The above two policies have been found by using the computer programme "Dynacode" (Hastings [4]) with semi-Markov option on B6700 computer.

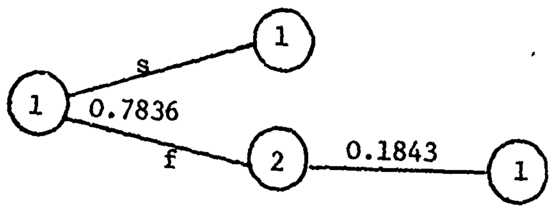
Using the algorithm developed in this paper we can find a better policy than the above. Let state 1 refer to the original (or replaced) equipment and state 2 refer to a repaired equipment. Then an action consists of taking the equipment from state i to state j, $i, j = 1, 2$ with a cost of $C_{MF}(i,j)$ on failure and $C_M(i,j)$ on survival. Consider the following network of an optimal policy:



In this network "f" represents failure and "s" represents survival. Initially we are at state 1. We have four sets of decisions to take. The first decision will be defined as the fourth stage decision and this decision is to replace (state 1) on survival to age 0.8237 and to repair (state 2) on earlier failure. The third stage decisions are to replace an equipment of state 2 at age 0.1918 or earlier failure, to replace an equipment of state 1 on survival to age 0.8066 and to repair it on an earlier failure. The second stage decisions are to replace an equipment of state 2 at age 0.1918 or earlier failure, to replace an equipment of state 1 on survival to age 0.8959 and to repair it on earlier failure. The first stage decision is to replace an equipment of state 2 at age 0.1918 or earlier failure and to replace an equipment of state 1 at age 0.5078 or earlier failure. After the first stage we come back to state 1 and the process starts all over again. The long-term average cost per unit time for this four stage policy is 336.80 which is cheaper than the optimal dynamic programming policy

for t=0.5 (cost = 344.00). It is also considerably cheaper than the age replacement policy (cost = 386.92).

By our algorithm we will find an n-stage optimal inefficient repair policy for a given value of n. Unlike dynamic programming, the age of planned action will be a decision variable in continuous domain and there will be no need to make t smaller and smaller to get a closer and closer approximation to the optimal policy in continuous time domain. In section 6 we have considered the problem of optimization over n. For the above example, the policy optimal over n, has the following structure,



with a long-term average cost per unit time of 326.01. It can be found by the method of section 6.2.

2. THE MODEL

2.1. Preliminary Definitions

Consider an equipment whose age at failure is a random variable $X > 0$. We assume that X possesses an absolutely continuous cumulative distribution function $F(\cdot)$ and density function $f(\cdot)$. We define the reliability at age x as,

$$R(x) = 1-F(x) , \tag{2.1}$$

so that $R(0) = 1$ and $R(\infty) = 0$. The failure rate at age x will be defined as

$$Z(x) = f(x)/R(x) . \tag{2.2}$$

We will assume that $Z(x)$ is a strictly increasing function of x so that the failure pattern of the equipment belongs to the strictly increasing failure rate (IFR) category. We will also assume that $Z(0) = 0$. Gamma and Weibull density with appropriate choice of the parameters satisfy

these assumptions.

Assume that just after a decision the age of the equipment is zero, the state of the equipment is i and the density of the age to failure is $f_i(\cdot)$. For the next decision we will divide the age of the equipment into intervals,

$$(T_0 = 0, T_1], (T_1, T_2], \ldots, (T_{m(i)-1}, T_{m(i)}]$$

such that if failure occurs in $(T_{r-1}, T_r]$ then action r is taken, $r=1', 2, \ldots, m(i)-1$. After this action the age of the equipment becomes zero and the state of the equipment becomes r. If failure occurs in $(T_{m(i)-1}, T_{m(i)}]$ or if the equipment survives to age $T_{m(i)}$ then action $m(i)$ is taken so that the equipment reaches age zero and state $m(i)$. All the density functions before or after actions belong to IFR category.

Starting with an equipment with age 0 and state 1 we will take such actions n times. A replacement will always bring back the equipment to state 1. At the last stage the only allowable action is replacement by an identical equipment (age 0, state 1) at age T_i or earlier failure. A cycle is now completed and the process starts all over again.

Let $U(i,j)$ represent a random variable denoting the actual time it takes for the action (i,j), moving the equipment from state i to state j to be completed. This random variable depends on (i,j) but not on the condition (viz failure or survival) of the equipment at the instant of action. Let

$$\mu(i,j) = E[U(i,j)] \tag{2.3}$$

The following costs will be considered,

(i) overhead cost of action on failure $C'_{MF}(i,j)$,

(ii) cost per unit time of action on failure $C''_{MF}(i,j)$,

(iii) the corresponding costs of action on survival $C'_M(i,j)$, $C''_M(i,j)$,

(iv) cumulative cost of operating the equipment from age 0 to age Y of action when the state of the equipment before action is i, $C_i Y^a$. (Scheaffer [13] has considered such costs). since we are dealing with IFR situations "ageing" will be present and we assume $a \geqslant 1$.

Hence the expected costs of actions are ,

$$\left.\begin{aligned} C_{MF}(i,j) &= C'_{MF}(i,j) + C''_{MF}(i,j)\mu(i,j), \text{ on failure} \\ C_M(i,j) &= C'_M(i,j) + C''_M(i,j)\mu(i,j), \text{ on survival.} \end{aligned}\right\} \quad (2.4)$$

We assume that expected cost on failure is greater than the expected cost on survival so that

$$\delta C_M(i,j) = C_{MF}(i,j) - C_M(i,j) > 0. \quad (2.5)$$

Our objective is to find

(i) the ages for the actions of the last stage,
(ii) the division points (number and values) $T_1, \ldots, T_{m(i)}$ of age and the corresponding actions, for each of $n-1$ stages,

such that the long-term average cost per unit time of the n stage process is a minimum.

It will be noticed that we have implicitly assumed that,

(i) there is no technological change in equipment design,
(ii) the costs are stationary over time,
(iii) the condition of the equipment is known with certainty so that no inspection costs need to be considered.

Assumptions (i) and (ii) give us a stationary model.

2.2 Long-Term Average Cost per Unit Time

We will consider an n-stage policy. At the beginning of the $(n-1)$th stage, let the state of the equipment be b . We will replace at age T_b or earlier failure.

Hence the expected operating cost is

$$C_b \int_0^{T_b} x^a f_b(x)dx + C_b T_b{}^a \int_{T_b}^{\infty} f_b(x)dx$$

$$= C_b a \int_0^{T_b} x^{a-1} R_b(x)dx \ .$$

The expected replacement cost is

$$C_{MF}(b,1)[1-R_b(T_b)] + C_M(b,1)R_b(T_b)$$

$$= C_{MF}(b,1) - \delta C_M(b,1)R_b(T_b)$$

Hence the expected cost is

$$N_1(b,1,T_b) = C_{MF}(b,1) - \delta C_M(b,1)R_b(T_b) + C_b a \int_0^{T_b} x^{a-1} R_b(x)dx \ . \qquad (2.6)$$

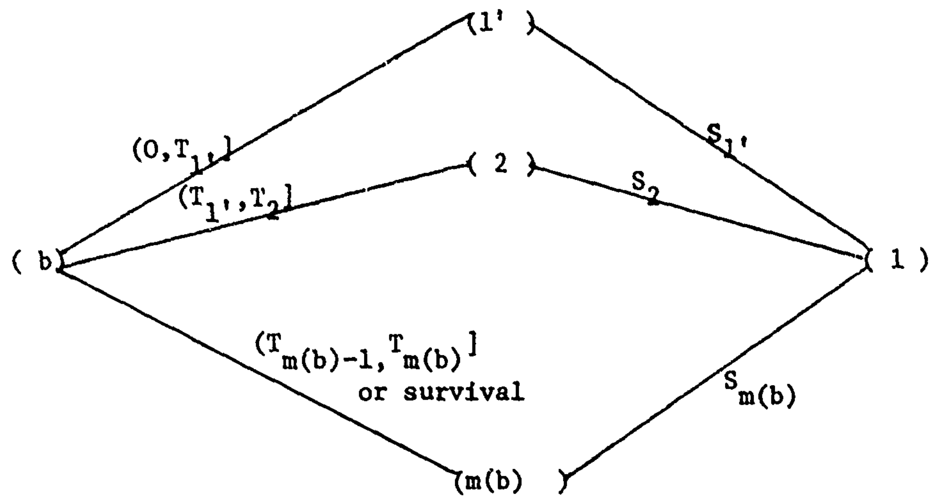
The symbol $N_1(\cdot,\cdot,\cdot)$ denotes that we are going from state b to state 1 at age T_b or earlier failure in one step.

The expected duration of this stage is

$$D_1(b,1,T_b) = \mu(b,1) + \int_0^{T_b} xf_b(x)dx + T_b \int_{T_b}^{\infty} f_b(x)dx$$

$$= \mu(b,1) + \int_0^{T_b} R_b(x)dx \ . \qquad (2.7)$$

Now suppose we have two stages to go. We are at state b . If a failure occurs in $(T_{r-1}, T_r]$ then action r will be taken, $r = 1, 2, \ldots, m(b)-1$ and $T_0=0$. If a failure occurs in $(T_{m(b)-1}, T_{m(b)}]$ or if the equipment survives up to $T_{m(b)}$ then action m(b) will be taken. Subsequently, actions appropriate to the nth stage will be taken. The network for the last two stages is as follows:



stage n-1 stage n

The T's represent the times at the $(n-1)$th stage and the S's represent the times at the nth stage.

Let,

$$I_1(b) = \text{set of actions at the stage next to } b$$
$$= (1', 2, \ldots, m(b)) . \tag{2.8}$$

$$T_1(b) = \text{set of times at the stage next to } b$$
$$= (T_1', T_2, \ldots, T_{m(b)}) . \tag{2.9}$$

$$S = \text{set of times at stage } n$$
$$= (S_{1'}, S_2, \ldots, S_{m(b)}) . \tag{2.10}$$

$$T_2(b) = \text{set of times at the two stages next to } b$$
$$= (T_1(b), S). \tag{2.11}$$

The expected operating cost for stage $n-1$

$$= C_b a \int_0^{T_{m(b)}} x^{a-1} R_b(x) dx ,$$

and the total expected cost for stages $n-1$ and n is

$$N_2(b,I_1(b),1,T_2(b))$$

$$= \sum_{r=1'}^{m(b)} [R_b(T_{r-1})-R_b(T_r)][C_{MF}(b,r) + N_1(r,1,S_r)]$$

$$+ R_b(T_{m(b)})[C_M(b, m(b)) + N_1(m(b), 1,S_{m(b)})]$$

$$+ C_b a \int_0^{T_{m(b)}} x^{a-1} R_b(x)dx \qquad (2.12)$$

where $N_2(\cdot,\cdot,\cdot,\cdot)$ represents the expected cost in going from b to 1 in two steps through intermediate states $I_1(b)$ and times $T_2(b)$.

Analogously, the total expected time for stages $n-1$ and n is

$$D_2(b,I_1(b), 1, T_2(b))$$

$$= \sum_{r=1'}^{m(b)} [R_b(T_{r-1}) - R_b(T_r)][\mu(b,r) + D_1(r,1,S_r)]$$

$$+ R_b(T_{m(b)})[\mu(b,m(b)) + D_1(m(b), 1, S_{m(b)})]$$

$$+ \int_0^{T_{m(b)}} R_b(x)dx . \qquad (2.13)$$

We can easily generalize (2.12) and (2.13) to cover all n stages.

We define a cycle to start with an equipment with age 0 and state 1 and to end when an equipment with age 0 and state 1 is next reached. The process now starts all over again. The n stage problem represents a cycle and the expected cycle cost is $N_n(1,I_{n-1}(1), 1, T_n(1))$ and the expected cycle time is $D_n(1,I_{n-1}(1), 1, T_n(1))$. Let $C(t)$ be the cost of operating this policy for time $(0,t)$. Then from renewal theory (Ross [9], Smith [14]), the long-term average cost per unit time is given by

$$C(I_{n-1}(1),T_n(1)) = \lim_{t\to\infty} \frac{C(t)}{t} = \frac{N_n(1,I_{n-1}(1),1,T_n(1))}{D_n(1,I_{n-1}(1),1,T_n(1))} . \qquad (2.14)$$

We want to find $I_{n-1}(1)$ and $T_n(1)$ such that $C(\cdot,\cdot)$ of (2.14) is a minimum.

3. OPTIMIZATION FOR A FIXED n

Let us define a function

$$F_r(b,I_{r-1}(b),1,T_r(b),\alpha) = N_r(b,I_{r-1}(b),1,T_r(b)) - \alpha\, D_r(b,I_{r-1}(b),1,T_r(b)), \quad (3.1)$$

for a given value of α. For b=1, fixed $I_{n-1}(1)$ and α we will minimize $F_n(\ldots)$ with respect to $T_n(1)$. We will show that for such a minimization a unique $\alpha=\alpha^* > 0$ exists such that

$$\min_{T_n(1)} F_n(\ldots) = 0, \text{ for } \alpha = \alpha^* ,$$

and,

$$\alpha^*(I_{n-1}(1)) = \min_{T_n(1)} C(I_{n-1}(1),T_n(1)) . \quad (3.2)$$

We will then find $\min\limits_{I_{n-1}(1)} \alpha^* (I_{n-1}(1))$. This technique has been used by Barlow and Proschan [2] and Brender [3]. We will use a mixture of calculus, dynamic programming and trial and error in our minimization.

3.1 Determination of α^*

Let,

$$F^*_n(b,I_{n-1}(b),1,T_n(I_{n-1}(b),\alpha),\alpha) = \min_{T_n(1)} F_n(b,I_{n-1}(b),1,T_n(b),\alpha) . \quad (3.3)$$

A useful property of $F^*_n(\cdots)$ is given in lemma 3.1 .

<u>Lemma 3.1</u>: The function $F^*_n(\cdots)$ of (3.3) is a strictly decreasing continuous function of α for fixed $I_{n-1}(\cdot)$.

Proof: The continuity follows immediately because $N_n(\cdots)$ and $D_n(\cdots)$ are functions with continuous derivatives and there is no division by zero involved in the minimization.

Let $\alpha_1 > \alpha_2$. Since $D_n(\cdots) > 0$,

$$F_n(b,I_{n-1}(b),1,T_n(b),\alpha_1) < F_n(b,I_{n-1}(b),1,T_n(b),\alpha_2)$$

and,

$$F_n^*(b,I_{n-1}(b),1,T_n(I_{n-1}(b),\alpha_1),\ \alpha_1)$$

$$< F_n^*(b,I_{n-1}(b),1,T_n(I_{n-1}(b),\alpha_2),\alpha_2),$$

by definition of $F_n^*(\cdots)$.

The inequality is strict and hence the lemma is proved.

Using this lemma we can prove lemma 3.2 below.

Lemma 3.2: There exists a unique $\alpha^*(I_{n-1}(1)) > 0$ such that $F_n^*(\cdots)$ of (3.3) is zero for $\alpha = \alpha^*(I_{n-1}(1))$ and $b=1$.

Proof: If we take actions on failures only then every element of $T_n(\cdot)$ tends to ∞ and then

$$F_n(1,I_{n-1}(1),1,\infty,\alpha)$$

$$= \sum_{r=1}^{n}[C_{MF}(j_{r-1},j_r) + C_b E(X^a_{j_{r-1}}) - \alpha\mu(j_{r-1},j_r) - \alpha E(X_{j_{r-1}})]$$

where the decision network for the n stage problem is as follows:

$$1 \xrightarrow[X_1]{f} j_1 \xrightarrow[X_{j_1}]{f} j_2 \xrightarrow[X_{j_2}]{f} \cdots \qquad \xrightarrow[X_{j_{n-2}}]{f} j_{n-1} \xrightarrow[X_{j_{n-1}}]{f} 1 .$$

Here "f" represents failure and $j_1, \ldots, j_{n-1}$ are the set of actions and X_{j_r} represents the age at failure

corresponding to the density function $f_{j_r}(\cdot)$. Also $j_0=1$, $j_n=1$.

If we choose α large enough but finite then $F_n(\cdots) < 0$. This policy is suboptimal for $F_n(\cdots)$, hence for the same α ,

$$F_n^*(1,I_{n-1}(1), T_n(I_{n-1}(1),\alpha),\alpha)< 0 .$$

For $\alpha = 0$,

$$F_n^*(1,I_{n-1}(1),1,T_n(I_{n-1}(1),0),0)$$

$$= \min_{T_n(1)} N_n(1,I_{n-1}(1),1,T_n(1))>0 ,$$

since $N_n(\cdots)>0$ for all $T_n(\cdot)>0$.

Since $F_n^*(\cdots)$ is a strictly decreasing continuous function of α (lemma 3.1), the lemma is proved.

We then notice from (2.14), (3.1), and (3.3) that

$$\frac{N_n(1,I_{n-1}(1),1,T_n(1))}{D_n(1,I_{n-1}(1),1,T_n(1))} > \alpha^*(I_{n-1}(1))$$

$$= \frac{N_n(1,I_{n-1}(1),1,T_n(I_{n-1}(1),\alpha^*))}{D_n(1,I_{n-1}(1),1,T_n(I_{n-1}(1),\alpha^*))}$$

$$\text{for all } T_n(1) . \tag{3.4}$$

Hence (3.2) holds, proving our assertion.

3.2 Dynamic Programming Formulation

By definitions of F,N and D

$$F_s(b, I_{s-1}(b),\ 1,\ T_s(b),\ \alpha)$$

$$= N_s(b, I_{s-1}(b), 1, T_s(b)) - \alpha D_s(b, I_{s-1}(b), 1, T_s(b))$$

$$\text{for,} \quad s = 1, 2, \ldots, n \qquad (3.5)$$

and,

$$F_{s+1}(b, I_s(b), 1, T_{s+1}(b), \alpha)$$

$$= \sum_{r=1'}^{m(b)} [R_b(T_{r-1}) - R_b(T_r)][C_{MF}(b,r) - \alpha\mu(b,r)$$
$$+ F_s(r, I_{s-1}(r), 1, T_s(r), \alpha)]$$
$$+ R_b(T_{m(b)})[C_M(b, m(b)) - \alpha\mu(b, m(b))$$
$$+ F_s(m(b), I_{s-1}(m(b)), 1, T_s(m(b)), \alpha)]$$
$$+ \int_0^{T_{m(b)}} [C_b a x^{a-1} - \alpha] R_b(x)\,dx\ ,$$

$$\text{for} \quad s = 1, 2, \ldots, n-1\ . \qquad (3.6)$$

This relationship is in the right form to apply the backward recurrence formula of dynamic programming. Backward recurrence have been discussed by Hastings [5] .

Let us define,

$$F_s^*(b, b', \alpha) = \min_{I_{s-1}(b), T_s(b)} F_s(b, I_{s-1}(b), b', T_s(b), \alpha)\ , \qquad (3.7)$$

$$\text{for} \quad s = 1, \ldots, n\ .$$

Then from (2.7) we get the dynamic programming recursion formula.

$$F^*_{s+1}(b,1,\alpha) = \min_{I_1(b),T_1(b)} \{ \sum_{r=1'}^{m(b)} [R_b(T_{r-1})-R_b(T_r)] \times$$

$$[C_{MF}(b,r)-\alpha\mu(b,r)+F^*_s(r,1,\alpha)]$$

$$+ R_b(T_{m(b)}) \; [C_M(b,m(b))-\alpha\mu(b,m(b))$$

$$+ F^*_s(m(b),1,\alpha)]$$

$$+ \int_0^{T_{m(b)}} [C_b a x^{a-1}-\alpha] R_b(x)dx\} ,$$

$$\text{for} \quad s = 1, \ldots, n . \qquad (3.8)$$

Let us define,

$$\left.\begin{aligned} A_r(b,\alpha) &= C_{MF}(b,r)-\alpha\mu(b,r)+F^*_s(r,1,\alpha) \\ & \qquad r = 1', \ldots, m(b) \\ A_{m(b)+1}(b,\alpha) &= C_M(b,m(b))-\alpha\mu(b,m(b))+F^*_s(m(b),1,\alpha) . \end{aligned}\right\} \qquad (3.9)$$

Then the right hand side of (3.8) is of the form

$$\min_{T_1(b)} [\Sigma_{r=1'}^{m(b)} (R_b(T_{r-1}) - R_b(T_r)) \min_{I_1(b)} A_r(b,\alpha)$$

$$+ R_b(T_{m(b)}) \min_{I_1(b)} A_{m(b)+1}(b,\alpha)$$

$$+ \int_0^{T_{m(b)}} (C_b a x^{a-1}-\alpha) R_b(x)dx] . \qquad (3.10)$$

Let us assume that the set of all feasible actions from state b are independent of $T_1(b)$ as to their feasibility. This statement means that, in the context of example 1.1, it is feasible to go from 1 to j , j = 1, 2 whatever $T_1(1)$ is and we have no restriction like "can go to 1 only if $T \le A$ and to 2 only if $A<T<\infty$" . In such an unrestricted case $\min_{I_1(b)} A_r(b,\alpha)$ is the same for $r=1', \ldots, m(b)$ but can be

...ent for $r=m(b)+1$. We can easily show this result for $F_1^*(b,1,\alpha)$. Hence by induction and backward recurrence of dynamic programming we have the following theorem:

Theorem 3.3: For any stage s and any state b suppose the set of immediate feasible actions $I_1(b)$ are not restricted by the immediate time values $T_1(b)$, then the optimal action is of the form "go to state i if failure occurs before age $T_{m(b)}$ and go to state j if the equipment survives to age $T_{m(b)}$". The values of i,j and $T_{m(b)}$ will depend on b and s.

Proof: As above.

By applying this theorem the type of policy applicable to our model becomes the same as mixed policy described in [11] and the algorithm discussed in section 2.3 of [11] apply with minor modification. The modifications required are due to non-zero operating cost.

Hence our algorithm consists of the following steps:

- (i) Find $N_s(\cdots)$ and $D_s(\cdots)$ for $s=1, \ldots, n$.
- (ii) Choose an α.
- (iii) Using (3.8) with $r=1'$, $T_0=0$ and $T_1=T_{m(b)}$, and (3.10) find $F_n^*(\cdots)$ for this α. Here we use theorem 3.3.
- (iv) If $F_n^*(\cdots)=0$ we have the optimal solution due to (3.4).
- (v) Using lemma 3.1 if $F_n^*(\cdots)>0$, increase α, or if $F_n^*(\cdots)<0$, decrease α.
- (vi) Go back to (iii).

We will illustrate the algorithm through an example.

Example 3.1: Let us consider an equipment having gamma distribution for the age to failure. Let

$$f_i(x)=\lambda_i^2 x e^{-\lambda_i x}, \quad \lambda_i>0, \quad x>0, \quad i=1,\ 2,\ 3.$$

This distribution belongs to the IFR family.

We can easily verify that the failure rate $Z_i(x)$ increases with λ_i. We assume,

	i	=	1	2	3
	λ_i	=	1	1.25	1.6
Hence,	$E(X_i)$	=	2	1.6	1.25

The cost and time data are as follows:

$$C''_{MF}(i,j)=100,\ C''_M(i,j)=50 \quad \text{for } i,j=1,2,3 .$$

$C'_{MF}(i,j)$ $C'_M(i,j)$ $\mu(i,j)$

$i \backslash j$	1	2	3	$i \backslash j$	1	2	3	$i \backslash j$	1	2	3
1	1050	630	430	1	75	50	40	1	0.5	0.4	0.2
2	1650	1120	790	2	175	135	95	2	0.5	0.3	0.1
3	2350	1580	990	3	375	290	195	3	0.5	0.2	0.1

Hence by equations (2.4) & (2.5)

$C_{MF}(i,j)$ $C_M(i,j)$ $\delta C_M(i,j)$

$i \backslash j$	1	2	3	$i \backslash j$	1	2	3	$i \backslash j$	1	2	3
1	1100	670	450	1	100	70	50	1	1000	600	400
2	1700	1150	800	2	200	150	100	2	1500	1000	700
3	2400	1600	1000	3	400	300	200	3	2000	1300	800

Also, a=1, and the operating costs are $C_1=20$, $C_2=30$, $C_3=40$.

We will consider up to 3 stage policies.

For one stage policy, the only possibility is as follows:

$$(1) \xrightarrow{\quad T \quad} (1)$$

$$F_1(1,1,T,\alpha)=C_{MF}(1,1)-\delta C_M(1,1)R_1(T)-\alpha\mu(1,1)+(C_1-\alpha)\int_0^T R_1(x)dx$$

Let $\alpha = 173.7014$. Then

$$F_1(1,1,T,\alpha)=1100-1000\,R_1(T)-0.5(173.7014)-(153.7014)\int_0^T R_1(x)dx.$$

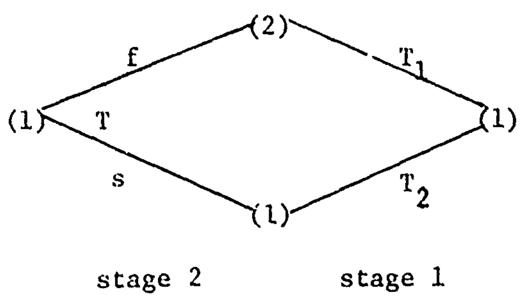
By differentiation optimal T is given by

$1000\, Z_1(T) - 153.7014 = 0$ i.e. $T = 0.181616$. This value is unique because we are considering only IFR situations,

and then, $F_1^*(1,1,\alpha) = -0.000159$.

Hence for one stage problem we replace at age 0.1816 or earlier failure and the long-term average cost per unit time is, $\alpha = 173.70$. This is the age replacement policy.

For the two stage problem $\alpha = 169.0557$ and the network for optimal policy is



f : failure before T

s : survival to T

$T = 0.273078$, $T_1 = 0.064083$, $T_2 = 0.175165$,

$$F_2^*(1,1,\alpha) = -0.000001 .$$

We will now show the working for a three stage problem.

At the last stage we can be at states $i=1,2,3$ and from there must go to the state 1 at age T_i, $i=1,2,3$ or earlier failure. Hence we must find $F_1^*(i,1,\alpha)$ for $i=1,2,3$. Let $\alpha = 167.6753$. Then

$$T_1 = 0.173262,\quad F_1^*(1,1,\alpha) = 4.077382$$

$$T_2 = 0.063396,\quad F_1^*(2,1,\alpha) = 111.911598$$

$$T_3 = 0.025973,\quad F_1^*(3,1,\alpha) = 314.527133 .$$

At the beginning of the second stage suppose we are at stage 1 . We want to find $F_2^*(1,1,\alpha)$. We will use theorem 3.3, equations (3.8), (3.9) and (3.10) .

j	$C_{MF}(1,j)-\alpha\mu(1,j)+F_1^*(j,1,\alpha)$	$C_M(1,j)-\alpha\mu(1,j)+F_1^*(j,1,a)$
1	1020.239732	20.239732*
2	714.841478*	114.841478
3	730.992073	330.992073

The '*' denotes the minimum in each column.

The F-value of going from 1 to 1 in two steps with transition to state 2 on failure before T and to state 1 on survival to T is given by,

$F_2(1,I_1(1),1,T_2(1),\ \alpha)$

$$= (1-R_1(T))(714.841478)+R_1(T)(20.239732)+(20-\alpha)\int_0^T R_1(x)dx$$

opt $T = 0.270009$, $F_2^*(1,1,\alpha) = \min_{T>0} F_2(\cdots) = 1.982496$.

We similarly find $F_2^*(2,1,\alpha)$ and $F_2^*(3,1,\alpha)$ and get

$F_2^*(2,1,\alpha) = 113.533769$, opt $T = 0.101584$, opt i on f is 3
opt j on s is 1

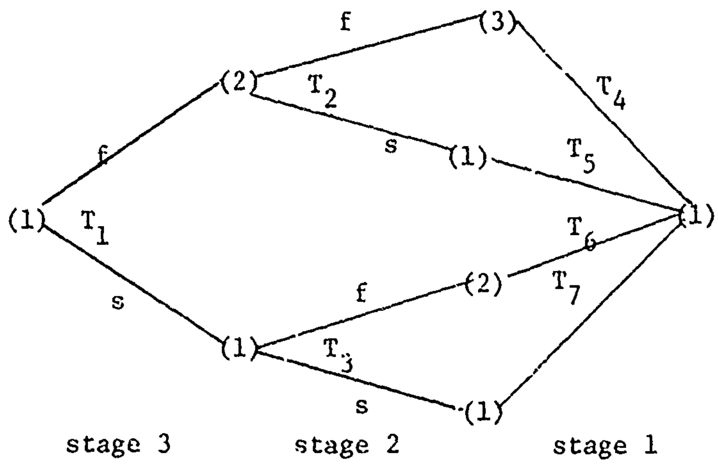
$F_2^*(3,1,\alpha) = 316.795952$, opt $T = 0.055555$, opt i on f is 3
opt j on s is 1 .

We now find $F_3^*(1,1,\alpha)$. As before,

j	$C_{MF}(1,j)-\alpha\mu(1,j)+F_2^*(j,1,\alpha)$	$C_M(1,j)-\alpha\mu(1,j)+F_2^*(j,1,\alpha)$
1	1018.144846	18.144846
2	716.463649*	116.463649
3	733.260892	333.260892

$F_3^*(1,1,\alpha) = -0.000059$, opt $T = 0.268187$.

Hence the solution network for the three stage problem is



$T_1 = 0.2682$, $T_2 = 0.1016$, $T_3 = 0.2700$, $T_4 = 0.0260$

$T_5 = 0.1733$, $T_6 = 0.0634$, $\alpha = 167.68$.

The α above is found by trial and error. We found for the two stage problem $\alpha = 169.0557$. We start with this value in the three stage problem. We find $F_3^*(1,1,\alpha)<0$. Hence we reduce α until we reach $F_3^*(1,1,\alpha)=0$ and find the corresponding optimal policy network as above.

We note that if we do not take inefficient repair into account then we always replace and our optimal policy is given by the solution of the one stage problem. For that problem $\alpha = 173.70$ and the three stage inefficient repair policy is 4% cheaper. In section 6 we will find the optimum policy for this problem over all stages.

4. DELAYED RENEWAL-REWARD PROCESS

So far in our models the process has always started with an equipment at state 1 and age 0 i.e. a new equipment. In many real life situations we may have to start the process with an old equipment, or we may find that replacement can be made only by an equipment different from the existing one when new. Assuming that there will be no further changes in the equipment design in the future we have what Ross [9] has called a "delayed renewal-reward process". Under this process the first cycle ends when the

equipment is replaced for the first time. All cycles from the second one onwards are identical and are of the type considered in sections 2 and 3 above. The first cycle is different from the rest. By a result analogous to theorem 3.16, pp. 52-54 of Ross [9] it can be easily shown that the long-term average cost per unit time for this delayed renewal-reward process is E(cycle cost)/E(cycle time) where the expected values refer to the identical cycles. (For details of proof, see Sarker [12], lemma 5.1). Hence from the second cycle onwards the theory developed so far applies without any modification. However, the optimal action for the first cycle has still to be found.

Suppose α^* denote the minimum long-term average cost per unit time from the second cycle onwards. Let $\alpha^* = N^*/D^*$. Suppose, in following a particular policy, N is the expected cost during the first cycle and D the corresponding expected time. Let,

$$N(\alpha^*) - \alpha^* D(\alpha^*) = \min\,[N - \alpha^* D] \,, \qquad (4.1)$$

minimization being over all stages, actions and times during the initial cycle. Then we have the following lemma:

<u>Lemma 4.1</u>: For large r, the ratio of expected cost over expected time for $r+1$ cycles will be minimized if we follow the policy given by $N(\alpha^*)-\alpha^*D(\alpha^*)$ for the first cycle and the policy given by α^* from the second cycle onwards.

<u>Proof</u>: We note that the cost ratio for $r+1$ cycles is

$$(N+rN^*)/(D+rD^*) \,,$$

where N and D refer to the initial cycle. From (4.1)

$$N - \alpha^* D \geq N(\alpha^*) - \alpha^* D(\alpha^*) \qquad (4.2)$$

for all N, D and $\alpha^* = N^*/D^*$.

Now,

$$(N+rN^*)/(D+rD^*) - [N(\alpha^*) + rN^*]/[D(\alpha^*) + rD^*]$$

$$=[\{ND(\alpha^*) - DN(\alpha^*)\} + rD^*\{(N-\alpha^*D) - (N(\alpha^*) - \alpha^*D(\alpha^*))\}]$$

$$\div (D+rD^*)(D(\alpha^*)+rD^*) \,, \text{ since } \alpha^* = N^*/D^*$$

≥ 0, for large r, by (4.2).

Hence the policy given by $N(\alpha^*) - \alpha^* D(\alpha^*)$ during the initial cycle and by α^* during the subsequent cycles minimizes the cost ratio over $r+1$ cycles for large r. Hence the lemma.

This lemma gives the policy during the initial cycle. If, however, in the initial cycle there exists a policy of not replacing ever and the cost ratio for this policy is less than α^*, then such a policy is optimal, rather than the policy given by the lemma.

5. RANDOM STATE/AGE REPLACEMENT POLICIES

Let us consider $C(.,.)$ of (2.14). In section 3 we have determined $I_{n-1}(1)$ and $T_n(1)$ through optimization. Let us now assume that these are determined by probability distributions. Barlow and Proschan (pp. 86-87 of [2]) have denoted such situations as random age replacement. We can easily prove the following lemma.

Lemma 5.1: Random age replacement policies are no cheaper than the optimal policies considered in section 3.

Proof: Similar to that of theorem 2.1, pp. 86-87 of [2]. For details see Sarkar [12], lemma 4.1.

6. OPTIMIZATION OVER n

In section 3 we have found the optimal policy for a given n. We will now consider the problem of minimizing $C(.,.)$ of (2.14) over positive integral values of n. For this purpose we will derive sufficient conditions for two situations.

6.1 Case 1

Suppose there exists a positive integer n_0 such that

$$\left.\begin{array}{l} F^*_{n_0}(1,1,\alpha^*) > 0 \text{ for } n=1,2,\ldots, n_0-1, n_0+1 \\ F^*_{n_0}(1,1,\alpha^*) = 0, \end{array}\right\} \quad (6.1)$$

then the sufficient conditions that this α^* is the minimum long-term average cost per unit time for all $n=1,2,3,\ldots$ are given by the following theorem:

Theorem 6.1: Let (6.1) hold. Suppose there exists a positive integer $q > 1$ such that

$$F_q^*(i,1,\alpha^*) > F_1^*(i,1,\alpha^*), \text{ for all feasible } i \neq 1$$

and,

$$F_{n_0}^*(1,1,\alpha^*) < F_{n_0+t}^*(1,1,\alpha^*), \; t=1, 2, \ldots, q-1 \; . \quad (6.3)$$

Then $F_{n_0}^*(1,1,\alpha^*)$ is minimum for all n .

To prove this theorem we need a preliminary lemma. An examination of (3.10) in conjunction with theorem 3.3 shows that the minimization for a fixed n requires minimization of functions of the form

$$G(T,A,B,\alpha) = [1-R(T)]A + R(T)B + \int_0^T [Cax^{a-1}-\alpha]R(x)dx \; ,$$

where A, B and α are independent of T .

Following the method of proof of lemma 3.1 we can easily show the following:

Lemma 6.2: (i) $\min\limits_{T>0} G(T,A,B,\alpha)$ is a strictly decreasing function of α for fixed A and B ;
(ii) $\min\limits_{T>0} G(T,A,B,\alpha)$ is an increasing function of $A[B]$ for fixed α and $B[A]$.

Proof: See Sarkar [11], lemma 2.3.1 for details.

Proof of theorem 6.1: For $n \geqslant n_0$ let us consider $F_{n+q-1}^*(1,1,\alpha^*)$ and compare it with $F_n^1(1,1,\alpha^*)$, where in $F_n^1(\ldots)$ we follow the same policy as that under $F_{n+q-1}^*(\ldots)$ up to stage $n-1$. Using (6.2) and applying lemma 6.2 repeatedly we get $F_{n+q-1}^*(1,1,\alpha^*) > F_n^1(1,1,\alpha^*)$.
This inequality is strict because it is strict for A in lemma 6.2 . Hence

$F_n^*(1,1,\alpha^*) < F_{n+q-1}^*(1,1,\alpha^*)$ for $n \geqslant n_0$

By repeated application of this inequality

$F^*_n(1,1,\alpha^*) < F^*_{n+sq-1}(1,1,\alpha^*)$, s=1, 2, ...

Using (6.3) we get,

$$F^*_{n_0}(1,1,\alpha^*) < F^*_{n_0+t}(1,1,\alpha^*) < F^*_{n_0+t+sq-1}(1,1,\alpha^*)$$

$$s = 1, 2, \ldots ; \quad t = 1, 2, \ldots , q-1 .$$

Hence,

$$F^*_{n_0}(1,1,\alpha^*) < F^*_n(1,1,\alpha^*) , \quad n > n_0 ,$$

showing that n_0 is optimal for all n . This proves the theorem .

Example 3.1 of Sarkar [11] gives an application of this result.

6.2 Case 2

If a finite n_0 does not exist then theorem 6.1 cannot be applied. Let $\alpha^*(n)$ be the minimum cost for an n-stage problem. In examples 1.1 and 3.1 $\alpha^*(n)$ decreases with n . In such situations we want to investigate whether $\alpha^*(n)$ tends to a finite limit as $n \to \infty$. If it does then we have a policy optimal over n .

The first step in the investigation relates to the conditions given in the following lemma:

Lemma 6.3: Suppose for some $n=n_0$ and some $\alpha > C_i$, we get

(i) $F^*_{n+1}(i,1,\alpha) - F^*_n(i,1,\alpha) = F^*(\alpha)$, independent of i ,

(ii) the optimal action at the (n+1)th stage is to go from i to j(i) on failure before age T(i) and to go from i to k(i) on survival to age T(i) ,

then the same $F^*(\alpha)$, j(i), k(i) and T(i) hold for all $n \geq n_0+1$.

Proof: From section 3.2 .

$$F^*_{n_0+1}(i,1,\alpha) = \min_T[(1-R_i(T))\min_j\{C_{MF}(i,j)-\alpha\mu(i,j)+F^*_{n_0}(j,1,\alpha)\}$$

$$+ R_i(1)\min_k\{C_M(i,k)-\alpha\mu(i,k)+F^*_{n_0}(k,1,\alpha)\}$$

$$+ \int_0^T\{C_i ax^{a-1}-\alpha\}R_i(x)dx]$$

where optimal values are $T=T(i)$, $j=j(i)$, $k=k(i)$.

Using condition (i) of the lemma

$$F^*_{n_0+2}(i,1,\alpha)=\min_T[(1-R_i(T))\min_j\{C_{MF}(i,j)-\alpha\mu(i,j)+F^*_{n_0}(j,1,\alpha)\}$$

$$+ R_i(T)\min_k\{C_M(i,k)-\alpha\mu(i,k)+F^*_{n_0}(k,1,\alpha)\}$$

$$+ \int_0^T\{C_i ax^{a-1}-\alpha\}R_i(x)dx]+ F^*(\alpha)$$

$$= F^*_{n_0+1}(i,1,\alpha) + F^*(\alpha)$$

and the optimal values for T, j and k do not change. Hence by induction, the lemma is proved.

Now suppose that we have found an α such that $F^*(\alpha)<0$ and $F^*(\alpha-\varepsilon)>0$ for every small $\varepsilon>0$. Hence $F^*_n(i,1,\alpha)$ is a strictly decreasing sequence and $F^*_n(i,1,\alpha-\varepsilon)$ is a strictly increasing sequence. So each must tend to a limit or drift to $\pm\infty$ as appropriate. Since $F^*_n(i,1,\beta)$ is a continuous function of β, $F^*_n(i,1,\alpha)$ can not tend to $-\infty$ and $F^*_n(i,1,\alpha-\varepsilon)$ can not tend to $+\infty$. Hence

$$\left.\begin{array}{l}\lim\limits_{n\to\infty} F^*_n(i,1,\beta) \text{ exists and is finite} \\ \text{for } \beta = \alpha \text{ and } \alpha-\varepsilon .\end{array}\right\} \qquad (6.4)$$

The policy corresponding to this limiting value of $F^*_n(\ldots)$ is the policy optimal over n .

We will apply these ideas to the data of example 3.1.

Example 6.1: Consider the data for example 3.1. We can easily verify that for $n=1, 2, \ldots, 7$, $\alpha^*(n)$ is decreasing in n. We further find that for $n=2, \ldots, 7$

i	$j(i)$	$k(i)$
1	2	1
2	3	1
3	3	1

Assume that $F_n^*(i,1,\alpha)$ tends to a finite limit for $i=1,2,3$ and $\lim_{n\to\infty} F_n^*(i,1,\alpha) = F^*(i,1,\alpha)$.

In the three equations

$$F_{n+1}^*(i,1,\alpha)=\min_T[(1-R_i(T))\{C_{MF}(i,j(i))-\alpha\mu(i,j(i))+F_n^*(j(i),1,\alpha)\}$$

$$+ R_i(T)\{C_M(i,k(i))-\alpha\mu(i,k(i))+F_n^*(k(i),1,\alpha)\}$$

$$- (\alpha-C_i)\int_0^T R_i(x)dx],\ i=1,2,3$$

we will substitute $F^*(i,1,\alpha) = F_n^*(i,1,\alpha)$ on both sides of the equations. The optimal α must be such that $F^*(1,1,\alpha) = 0$. We get by trial and error (correct to the places of decimals shown).

i	$F^*(i,1,\alpha)$	$T(i)$	
1	0	0.262522	
2	111.053303	0.099098	$\alpha=165.0465$
3	314.186798	0.054150	

We note that these solutions are unique for α, $T(i)$, $i=1,2,3$ but not for $F^*(i,1,\alpha)$. In fact, the solutions are unique for

$$F^*(2,1,\alpha) - F^*(1,1,\alpha)\ ,\quad F^*(3,1,\alpha) - F^*(1,1,\alpha).$$

With these values we will verify whether lemma 6.3 is applicable and condition (6.4) holds. Numerical calculations give us the following:

α	$F^*(\alpha)$	$n_0(\alpha)$
164.0464	0.000070	6
164.0465	-0.000005	6

The values of $j(i)$, $k(i)$ and $T(i)$, $i=1,2,3$ remain as given above.

Hence lemma 6.3 is applicable and condition (6.4) holds. The policy optimal over n is then

state 1: go to state 2 on failure before $T(1) = 0.2625$
go to state 1 on survival to $T(1)$

state 2: go to state 3 on failure before $T(2) = 0.0991$
go to state 1 on survival to $T(2)$

state 3: go to state 3 on failure before $T(3) = 0.0542$
go to state 1 on survival to $T(3)$.

The long term average cost per unit time is $\alpha = 165.05$.

REFERENCES

[1] Barlow, R.E. and Proschan, F., Planned Replacement, Ch. 4, STUDIES IN APPLIED PROBABILITY AND MANAGEMENT SCIENCE, Arrow, K., Karlin, S. and Scarf, H., (ed.), Stanford University Press, Stanford, 1962

[2] Barlow, R.E. and Proschan, F., MATHEMATICAL THEORY OF RELIABILITY, Wiley, New York, 1965

[3] Brender, D.M., A Surveillance Model for Recurrent Events, IBM Watson Research Centre Report RC-837, Yorktown Heights, New York, 1962

[4] Hastings, N.A.J., DYNACODE HANDBOOK, Third Ed., Albany Interactive, P.O. Box 79, Ferntree Gully, Victoria 3156, Australia, 1978

[5] Hastings, N.A.J., DYNAMIC PROGRAMMING WITH MANAGEMENT APPLICATIONS, Butterworths, London, 1973

[6] Howard, R.A., DYNAMIC PROGRAMMING AND MARKOV PROCESSES, Wiley, New York, 1960

[7] Jardine, A.K.S. (ed.), OPERATIONAL RESEARCH IN MAINTENANCE, Manchester University Press, Manchester, 1970

[8] Jardine, A.K.S., MAINTENANCE REPLACEMENT AND RELIABILITY, Pitman, London, 1973

[9] Ross, S., APPLIED PROBABILITY MODELS WITH OPTIMIZATION APPLICATIONS, Holden-Day, San Francisco, 1970

[10] Sarkar, T.K., Age Replacement with Inefficient Repair, PROCEEDINGS OF THE 3RD NATIONAL CONFERENCE OF THE AUSTRALIAN SOCIETY FOR OPERATIONS RESEARCH, Adelaide, 312-342, 1977

[11] Sarkar, T.K., Optimal Replacement Policy Under Inefficient Repair with Multiple Repair Options, Working Paper No. 18, Department of Econometrics and Operations Research, Monash University, Australia, Nov. 1977

[12] Sarkar, T.K., Inefficient Repair Under Age/Time Dependent Costs, Working Paper No. 28, Department of Econometrics and Operations Research, Monash University, Australia, June 1978

[13] Scheaffer, R.L., Optimum Age Replacement Policies with an Increasing Cost Factor, TECHNOMETRICS, 13, 139-144, 1971

[14] Smith, W.L., Renewal Theory and its Ramifications, J. ROY, STATIST. SOC., series B, 20, 243-302, 1958

[15] Tahara, A. and Nishida, T., Optimal Replacement Policy for Minimal Repair Model, J. of O.R., SOC. OF JAPAN, 18, 113-124, 1975

[16] White, D.J., DYNAMIC PROGRAMMING, Oliver & Boyd, London, 1969.

OPTIMUM INSPECTION-ORDERING POLICIES
WITH
TWO TYPES OF LEAD TIMES

NAOTO KAIO and SHUNJI OSAKI

Department of Industrial Engineering,
Hiroshima University,
Hiroshima 730, Japan

ABSTRACT. In this paper we consider an inspection policy of a one-unit system with two types of lead times and costs for an inspection, orders, and a failure, where the lead time for an expedited order is not greater than that for a regular one. We call this policy inspection-ordering policy, discuss the optimum inspection-ordering policy maximizing a cost effectiveness, and obtain a finite and unique optimum inspection-ordering time and its lower limit under certain conditions. Finally, we present the numerical examples assumming a Gamma distribution.

1. INTRODUCTION

In recent years, systems have been complicated and large-scale, namely cars, airplanes, plants, etc.. As a result of this, the reliability, especially the maintenability engineering has developped. In that field, many papers concerning replacement policies have been discussed (e.g., see Barlow et al. [3-7, 9]). Especially, an age and a block replacement policy and an inspection policy discussed by Barlow and Proschan [3] are well-known. In these models, spares are always provided if necessarily, i.e., they are always on hand. However, in practice it is natural that spares are not always on hand and there exists any kind of delay between a failure of a system and a replacement (e.g., see Allen et al. [1, 2, 10]). If we consider this kind of delay, the above-mentioned models by Barlow et al. become included in a class of ordering policies. In this paper we discuss the inspection policy by Barlow and Proschan (see [3, p. 107]) with the above delay, where this policy is named an <u>inspection-ordering policy</u>.

In this paper we treat the extended inspection policy in the above sense, and our purpose is to seek an optimum inspection-ordering policy. We illustrate this policy in brief: At a prespecified time a system is inspected. Then if a failure of the system is detected, an order of a new unit is made expeditiously, which is called an <u>expedited order</u>. Reversely, if the system is still operating, the order of the unit is made regularly, which is called a <u>regular order</u>. After a lead time, the ordered unit is delivered and the system is replaced by the unit immediately, irrespective of the situation that the system is operating or not. This model is suitable to the real system. Of course, if we neglect the lead time, i.e., the above delay does not exist, this policy becomes the well-known "inspection policy" by Barlow and Proschan.

Especially, in this paper we discuss the optimum inspection-ordering policies with two types of lead times and four linear costs. That is, we introduce two types of lead times for a regular order and an expedited order, and four costs for an inspection, a regular order, an expedited order, and a failure. Of course, we assume that the lead time for an expedited order is not greater than that for a regular one. Because, in the case that we have detected the system failure by an inspection, it might be plausible that we tend to obtain the new unit in shorter time by the expedited order. The analysis is done as follows: Introducing the above two types of lead times and four costs and noticing that every replacement time

instant is a regeneration point, we derive the steady-state availability, the expected cost per unit time in the steady-state, and the cost effectiveness generated by the above two quantities. We seek the optimum inspection-ordering policies maximizing the cost effectiveness using the failure rates of the operating unit. It is shown that there exists a unique and finite optimum policy under certain conditions. It is further shown that there exists a lower limit of such an optimum policy. Fanally, the numerical examples of the optimum inspection-ordering policies in this paper are presented assumming that the failure time distribution of the operating unit is a Gamma distribution.

2. MODEL AND ASSUMPTIONS

We cosider a one-unit system, where each failed unit is scrapped and not repaired and each spare is only provided after a lead time by an order. We assume the following: The planning horizon is infinite. The original unit begins operating at time 0 and is first inspected at a prespecified time instant $t_0 \in [0, \infty)$. The inspection is done instantaneously and cause neither the failure nor the deterioration of the unit. Consequently, if the system has failed, the expedited order for the spare is made immediately, and after an expedited lead time L_1 the spare takes over operation as soon as it is delivered. On the other hand, if the system is still operating, the regular order is made immediately. After a regular lead time L_2 the spare is delivered, and the original unit is replaced and taken over operation by the delivered spare immediately, irrespective of the situation of the original unit. Subsequently, cycles are repeated in a similar fashion.

We assume that the failure time for each unit obeys an identical and arbitrary distribution $F(t)$ with a finite mean $1/\lambda$ and a density $f(t)$. The costs considered here are the following; a constant cost k_1 per unit time is suffered for the system failure, and costs c_1, c_2, and c_3 are suffered for the expedited order, the regular order, and the inspection, respectively, which are made at the time instant t_0. We assume that $L_1 \leq L_2$, $c_1 > c_2$ and $c_1 + k_1 L_1 > c_2 + k_1 L_2$. These assumptions are reasonable.

Under the above-mentioned assumptions, we define a period from the beginning of the original unit to replacement, i.e., from one replacement to next replacement as

one cycle, and discuss the optimum inspection-ordering policies.

3. ANALYSIS AND THEOREMS

The mean time of one cycle is

$$(t_0 + L_1)F(t_0) + (t_0 + L_2)\overline{F}(t_0)$$

$$= t_0 + L_1F(t_0) + L_2\overline{F}(t_0) , \quad (1)$$

where $\overline{F}(t) \equiv 1 - F(t)$. In this mean time, the mean time of operation is

$$\int_0^{t_0} tdF(t) + \int_{t_0}^{t_0+L_2} tdF(t) + \int_{t_0+L_2}^{\infty} (t_0 + L_2)dF(t)$$

$$= \int_0^{t_0+L_2} \overline{F}(t)dt . \quad (2)$$

Thus, the steady-state availability becomes

$$A(t_0) = \int_0^{t_0+L_2} \overline{F}(t)dt \Big/ [t_0 + L_1F(t_0) + L_2\overline{F}(t_0)] , \quad (3)$$

(see Ross [8, p. 95]) .

Next, the following two expected costs are considered: (i) When the original unit has failed, the state of the system failure continues until the spare is delivered. The expected cost during that period is

$$k_1[\int_0^{t_0} (t_0 + L_1 - t)dF(t) + \int_{t_0}^{t_0+L_2} (t_0 + L_2 - t)dF(t)]$$

$$= k_1[(L_1 - L_2)F(t_0) + \int_0^{t_0+L_2} F(t)dt] . \quad (4)$$

(ii) The expected inspect-order cost is

$$(c_3 + c_1)F(t_0) + (c_3 + c_2)\overline{F}(t_0)$$

$$= c_1F(t_0) + c_2\overline{F}(t_0) + c_3 . \quad (5)$$

Thus, the expected cost per unit time in the steady-state is

$$K(t_0) = [k_1\{(L_1 - L_2)F(t_0) + \int_0^{t_0+L_2} F(t)dt\} + c_1F(t_0) + c_2\overline{F}(t_0) + c_3]/[t_0 + L_1F(t_0) + L_2\overline{F}(t_0)] , \quad (6)$$

(see Ross [8, p. 52]).

Now, we define a cost effectiveness as [System effectiveness]/[Cost], especially, [The steady-state availability]/[The expected cost per unit time in the steady-state] (e.g., see Zelen [11]). Then, the cost effectiveness of this model is

$$E(t_0) = \int_0^{t_0+L_2} \overline{F}(t)dt/[k_1\{\int_0^{t_0+L_2} F(t)dt + (L_1 - L_2) \times F(t_0)\} + c_1F(t_0) + c_2F(t_0) + c_3] , \qquad (7)$$

and clearly $E(t_0) > 0$, $E(t_0) \to 0$ as $t_0 \to \infty$, and

$$E(0) = \int_0^{L_2} \overline{F}(t)dt/[k_1\int_0^{L_2} F(t)dt + c_2 + c_3] . \qquad (8)$$

Define the numerator of the derivative of the right-hand side in (7) as

$$\begin{aligned} e(t_0) = {} & [1 - R(t_0)][k_1\{(L_1 - L_2)F(t_0) + \int_0^{t_0+L_2} F(t)dt\} \\ & + c_1F(t_0) + c_2\overline{F}(t_0) + c_3] - \int_0^{t_0+L_2} \overline{F}(t)dt \\ & \times[\{k_1(L_1 - L_2) + (c_1 - c_2)\}r(t_0) + k_1\{R(t_0) \\ & + F(t_0)/\overline{F}(t_0)\}] , \end{aligned} \qquad (9)$$

where $R(t_0) \equiv [F(t_0 + L_2) - F(t_0)]/\overline{F}(t_0)$, $r(t_0) \equiv f(t_0)$ $\overline{F}(t_0)$. Both functions $R(t_0)$ and $r(t_0)$, which are called <u>failure rates</u>, are assumed to be differentiable and have the same monotone properties (see Barlow and Proschan [3, p. 23]). Further, $e(\infty) < 0$ and

$$\begin{aligned} e(0) = {} & [1 - R(0)][k_1\int_0^{L_2} F(t)dt + c_2 + c_3] - \int_0^{L_2} \overline{F}(t)dt \\ & \times[\{k_1(L_1 - L_2) + (c_1 - c_2)\}r(0) + k_1R(0)] . \end{aligned} \qquad (10)$$

Here, we have the following theorem, which shows the sufficient condition for the existence of the optimum policy.

[Theorem 1] If $e(0) > 0$, then there exists at least an optimum inspection-ordering time t_0^* $(0 < t_0^* < \infty)$ maximizing the cost effectiveness $E(t_0)$.

[Proof] By differentiating $\log E(t_0)$ with respect to t_0, we have

$$[d \log E(t_0)]/[dt_0]$$

$$= \overline{F}(t_0) \; [\; [1 - R(t_0)]/[\int_0^{t_0+L_2} \overline{F}(t)dt] - [\{k_1(L_1 - L_2)$$

$$+ (c_1 - c_2)\}r(t_0) + k_1\{R(t_0) + F(t_0)/\overline{F}(t_0)\}]$$

$$/[k_1\{(L_1 - L_2)F(t_0) + \int_0^{t_0+L_2} F(t)dt\} + c_1F(t_0)$$

$$+ c_2\overline{F}(t_0) + c_3]\;] \quad . \tag{11}$$

For large t_0, $[d \log E(t_0)]/[dt_0] < 0$, and for small t_0,

$$[d \log E(t_0)]/[dt_0]$$

$$\simeq \overline{F}(t_0) \; [\; [1 - R(0)]/[\int_0^{L_2} \overline{F}(t)dt] - [\{k_1(L_1 - L_2)$$

$$+ (c_1 - c_2)\}r(0) + k_1R(0)]/[k_1\int_0^{L_2} F(t)dt$$

$$+ c_2 + c_3]\;] \quad , \tag{12}$$

which implies that, if the bracket of the right-hand side is positive, i.e., $e(0) > 0$, then there exists at least an optimum inspection-ordering time t_0^* $(0 < t_0^* < \infty)$ maximizing the cost effectiveness $E(t_0)$. Q.E.D.

The above theorem shows that there exists at least an optimum inspection-ordering time t_0^*, which is not necessarily unique, under a certain condition. However, supposing the monotone properties of the failure rate, especially the strictly increasing property, we have the following main theorem.

[Theorem 2] Suppose that the failure rate is strictly increasing.

(i) If $e(0) > 0$, then there exists a finite and unique optimum inspection-ordering time t_0^* $(0 < t_0^* < \infty)$ satisfying $e(t_0) = 0$, and the cost effectiveness is

$$E(t_0^*) = [1 - R(t_0^*)]/[\{k_1(L_1 - L_2) + (c_1 - c_2)\}r(t_0^*)$$

$$+ k_1\{R(t_0^*) + F(t_0^*)/\overline{F}(t_0^*)\}] \; . \tag{13}$$

(ii) If $e(0) \leq 0$, then the optimum inspection-ordering time is $t_0^* = 0$, i.e., an inspect-order for a system is made at the same time instant as the beginning of the original unit, and the cost effectiveness is given in (8).

[Proof] By differentiating $E(t_0)$ with respect to t_0 and setting it equal to zero, we have the equation

$e(t_0) = 0$. Further, we have

$$e'(t_0) = -R'(t_0)[k_1\{(L_1 - L_2)F(t_0) + \int_0^{t_0+L_2} F(t)dt\} + c_1F(t_0) + c_2\overline{F}(t_0) + c_3] - \int_0^{t_0+L_2} \overline{F}(t)dt \times[\{k_1(L_1 - L_2) + (c_1 - c_2)\}r'(t_0) + k_1\{R'(t_0) + r(t_0)/\overline{F}(t_0)\}] . \quad (14)$$

We treat the case that the failure rate is strictly increasing. Thus, we have that $e'(t_0) < 0$, i.e., $e(t_0)$ is strictly decreasing.

If $e(0) > 0$, then there exists a finite and unique t_0^* $(0 < t_0^* < \infty)$ which maximizes the cost effectiveness $E(t_0)$ as a finite and unique solution to $e(t_0) = 0$, since $e(\infty) < 0$ and $e(t_0)$ is strictly decreasing and continuous. Substituting the relation of $e(t_0^*) = 0$ into $E(t_0^*)$ in (7) yields (13).

If $e(0) \leq 0$, then for any non-negative t_0, $E'(t_0) \leq 0$ and thus $E(t_0)$ is a strictly decreasing function. Thus, the optimum inspection-ordering time is $t_0^* = 0$. Q.E.D.

Moreover, in case (i) in Theorem 2, we can give a lower limit for the optimum inspection-ordering time t_0^*, which might be useful for numerical computation.

[Theorem 3] Suppose that the failure rate is strictly increasing and $e(0) > 0$. If $\overline{t}_0$ is a solution satisfying the equation $h(t_0) = 0$, there exists a unique $\overline{t}_0$ such that $\overline{t}_0 < t_0^*$, where

$$h(t_0) = [1 - R(t_0)][k_1\int_0^{L_2} F(t)dt + c_2] - \int_0^{t_0+L_2} \overline{F}(t)dt[\{k_1(L_1 - L_2) + (c_1 - c_2)\}r(t_0) + k_1\{R(t_0) + F(t_0)/\overline{F}(t_0)\}]. \quad (15)$$

[Proof] We have

$$e(t_0) - h(t_0) = [1 - R(t_0)][k_1\int_{L_2}^{t_0+L_2} \bar{F}(t)dt + \{k_1(L_1 - L_2) + (c_1 - c_2)\}F(t_0) + c_3] . \quad (16)$$

Thus, we obtain the inequality $h(t_0) < e(t_0)$. Thus, if there exists a solution satisfying $h(t_0) = 0$, then the solution $\bar{t}_0$ is a unique one and $\bar{t}_0 < t_0^*$, since $h(t_0)$ is strictly decreasing. Q.E.D.

4. NUMERICAL EXAMPLES

In the preceding sections, we have shown that there exists the optimum inspection-ordering time t_0^* maximizing the cost effectiveness $E(t_0)$ under certain conditions. In this section, we will present the numerical examples of the optimum inspection-ordering time t_0^* and its associated quantities, $E(t_0^*)$, $A(t_0^*)$, $K(t_0^*)$, etc., in particular, the dependence of the regular lead time L_2 for illustration.

Assume that the failure time obeys a Gamma distribution with a shape parameter 2, i.e.,

$$F(t) = 1 - (1 + \alpha t)\exp(-\alpha t) \quad (\alpha > 0) .$$

Then,

$$\bar{F}(t) = (1 + \alpha t)\exp(-\alpha t) ,$$
$$f(t) = \alpha(\alpha t)\exp(-\alpha t) ,$$
$$r(t) = \alpha^2 t/(1 + \alpha t) ,$$
$$r'(t) = \alpha^2/(1 + \alpha t)^2 ,$$
$$R(t) = 1 - e^{-\alpha L_2} - \alpha L_2 e^{-\alpha L_2}/(1 + \alpha t) ,$$
$$R'(t) = \alpha^2 L_2 e^{-\alpha L_2}/(1 + \alpha t)^2 ,$$

and

$$1/\lambda = 2/\alpha .$$

Thus, this distribution has a strictly increasing failure rate with $r(0) = 0$ and $r(\infty) = \alpha$.

Under these assumptions, we obtain the optimum inspection-ordering time t_0^* and its associated cost

effectiveness $E(t_0^*)$, the steady-state availability $A(t_0^*)$, the expected cost per unit time in the steady-state $K(t_0^*)$, and the percent gain $[(E(t_0^*) - E(0))/E(0)]\times 100(\%)$ for $E(0)$ using Theorem 2, where a Newton-Raphson method is applied. Numerical examples are presented in Table I as the dependence of the regular lead time L_2, where all the other parameters are fixed.

5. CONCLUDING REMARKS

We have obtained three theorems on the optimum inspection-ordering policy maximizing the cost effectiveness, which is defined as [The steady-state availability] /[The expected cost per unit time in the steady-state], and we have shown that there exist a finite and unique optimum inspection-ordering policy and its lower limit under certain conditions, introducing two types of lead times for the expedited order and the regular one and four costs for the inspection, the orders, and the failure. The numerical examples have been presented supposing a Gamma distribution with a shape parameter 2, which show the dependence of the regular lead time in the optimum inspection-ordering time and its associated quantities.

Furthermore, in this model we can obtain some interesting results, even if we suppose the decreasing failure rate or we use any non-linear costs, e.g., $C_i(t_0)$ $(i = 1, 2, 3)$ instead of c_i.

Table I. Dependence of the regular lead time L_2 in the optimum inspection-ordering time t_0^* and its associated cost effectiveness $E(t_0^*)$, the steady-state availability $A(t_0^*)$, the expected cost per unit time in the steady-state $K(t_0^*)$, and the percent gain $[(E(t_0^*) - E(0))/E(0)] \times 100(\%)$ for $E(0)$ ($F(t) = 1 - (1 + \alpha t)\exp(-\alpha t)$, $1/\lambda = 2/\alpha = 50$, $c_1 = 3$, $c_2 = 1$, $c_3 = 0.5$, $k_1 = 0.05$, $L_1 = 5$)

L_2	t_0^*	$E(t_0^*)$	$A(t_0^*)$	$K(t_0^*)$	$\frac{E(t_0^*) - E(0)}{E(0)} \times 100(\%)$
5	14.80	10.16	0.9285	0.0914	207.01
6	13.60	10.29	0.9347	0.0909	159.84
7	12.47	10.42	0.9392	0.0901	126.46
8	11.39	10.56	0.9422	0.0892	101.70
9	10.37	10.71	0.9438	0.0882	82.70
10	9.41	10.86	0.9443	0.0870	67.74
11	8.50	11.02	0.9437	0.0857	55.73
12	7.64	11.18	0.9423	0.0843	45.93
13	6.83	11.36	0.9402	0.0828	37.84
14	6.06	11.53	0.9375	0.0813	31.10
15	5.33	11.72	0.9345	0.0797	25.44
16	4.64	11.91	0.9311	0.0782	20.66
17	4.00	12.10	0.9275	0.0766	16.61
18	3.39	12.30	0.9237	0.0751	13.18
19	2.82	12.50	0.9199	0.0736	10.26
20	2.28	12.71	0.9159	0.0721	7.79
21	1.78	12.92	0.9120	0.0706	5.69
22	1.31	13.13	0.9080	0.0691	3.93
23	0.87	13.35	0.9040	0.0677	2.45
24	0.47	13.57	0.8999	0.0663	1.23
25	0.09	13.78	0.8958	0.0650	0.23
26	0.00	14.08	0.8899	0.0632	0.00

REFERENCES

[1] Allen, S.G. and D.A. D'Esopo, AN ORDERING POLICY FOR REPAIRABLE STOCK ITEMS, Opns. Res., Vol 16, No 3, pp. 669-674, 1968

[2] Allen, S.G. and D.A. D'Esopo, AN ORDERING POLICY FOR STOCK ITEMS WHEN DELIVERY CAN BE EXPEDITED, Opns. Res., Vol 16, No 4, pp. 880-883, 1968

[3] Barlow, R.E. and F. Proschan, MATHEMATICAL THEORY OF RELIABILITY, Wiley, New York, 1965

[4] Barlow, R.E. and F. Proschan, STATISTICAL THEORY OF RELIABILITY AND LIFE TESTING: PROBABILITY MODELS, Holt, Rinehart and Winston, New York, 1975

[5] Cleroux, R. and M. Hanscom, AGE REPLACEMENT WITH ADJUSTMENT AND DEPRECIATION COSTS AND INTEREST CHARGES, Technometrics, Vol 16, No 2, pp. 235-239, 1974

[6] McCall, J.J., MAINTENANCE POLICIES FOR STOCHASTICALLY FAILING EQUIPMENT: A SURVEY, Mgmt. Sci., Vol 11, No 5, pp. 493-524, 1965

[7] Osaki, S. and T. Nakagawa, A NOTE ON AGE REPLACEMENT, IEEE Trans. Reliability, Vol R-24, No 1, pp. 92-94, 1975

[8] Ross, S.M., APPLIED PROBABILITY MODELS WITH OPTIMIZATION APPLICATIONS, Holden-Day, San Francisco, 1970

[9] Scheaffer, R.L., OPTIMUM AGE REPLACEMENT POLICIES WITH AN INCREASING COST FACTOR, Technometrics, Vol 13, No 1, pp. 139-144, 1971

[10] Wiggins, A.D., A MINIMUM COST MODEL OF SPARE PARTS INVENTORY CONTROL, Technometrics, Vol 9, No 4, pp. 661-665, 1967

[11] Zelen, M. (ed.), STATISTICAL THEORY OF RELIABILITY, The University of Wisconsin Press, Madison, 1964

A CHANCE-CONSTRAINED ZERO-ONE PROGRAMMING MODEL FOR RESOURCE CONSTRAINED CAPITAL BUDGETING

B.N. LOHANI

Environmental Engineering Division
Asian Institute of Technology
P.O. Box 2754, Bangkok
Thailand

For Presentation

At

The Pacific Conference On
OPERATIONS RESEARCH

April 23-28, 1979

Seoul, Korea

A CHANCE-CONSTRAINED ZERO-ONE PROGRAMMING MODEL FOR RESOURCE CONSTRAINED CAPITAL BUDGETING

B.N. LOHANI

Environmental Engineering Division
Asian Institute of Technology
P.O. Box 2754, Bangkok

ABSTRACT. In this paper a probabilistic programming model (Chance-constrained zero-one programming) has been developed for resource constrained capital budgeting problems. The model can be used to solve decision making problems in which one seeks to choose an optimal combination from a large number of projects so that the overall profit is maximized. The probabilistic programming model could not be solved in the form in which it was developed; however the use of the linear decision rule allowed the model to convert into an equivalent zero-one integer programming model which could be easily solved by the available algorithms.

1. INTRODUCTION

There are many possible formulation in which an optimization model for resource constrained capital budgeting problem can be formulated. The type of the formulation obviously depends upon the objective and the nature of the resources constrained. Deterministic formulation for the resource constrained capital budgeting problem is common. However, where the resources available are stochastic, it is more logical to formulate a probabilistic programming model. In this paper, a probabilistic formulation for the resource constrained capital budgeting problem; with the objective to choose an optimal combination of problems from a large number of projects so that the profit is maximized, has been presented. The model considers resource as stochastic parameters and provides a solution technique.

2. FORMULATION OF THE PROBLEM

A deterministic formulation of the resource constrained capital budgeting problem has been reported elsewhere [1]. In the same context, the probabilistic formulation of the optimization problem is presented herein.

2.1. Objective Function

The objective is to maximize the profit of the total projects (n) where there are k different policies for each project. Therefore the objective function is

$$\text{Maximize} \sum_{i=1}^{n} \sum_{j=1}^{k} C_{ij} \; X_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where

C_{ij} = profit if the ith project is implemented with jth policy

X_{ij} = 0 or 1; 0 if the ith project should be implemented with jth policy and 1 if the ith policy should not be implemented with jth policy.

2.2. Constraints

The constraints specify that the resources needed to implement ith project should be less than the resource

available. As the resource available is stochastic in nature, a chance constrained specifying the probability of meeting the constraint can be formulated. That is

$$\text{Prob}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ijp} X_{ij} \leq R_p\right] \geq \alpha_p, \; p = 1, 2 \ldots\ldots m$$

$$\sum_{j=1}^{k} X_{ij} \leq 1 \qquad i=1, 2\ldots\ldots\ldots n$$

$$X_{ij} = 0, 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where

a_{ijp} = amount of ith resource needed to implement project i with policy j

R_p = amount of ith resource available, a stochastic variable with known probability density function

α_p = degree of certainty of meeting the constraint

Prob. (P) = Probability

2.3. The Model

The chance constrained zero one programming model(CCZOC) for capital budgeting can hence be written as

$$\text{Maximize} \sum_{i=1}^{n}\sum_{j=1}^{k} C_{ij} X_{ij}$$

Subject to

$$P\left[\sum_{j=1}^{n}\sum_{j=1}^{k} a_{ijp} X_{ij} \leq R_p\right] \geq \alpha_p, \; p = 1, 2, \ldots m \quad (3)$$

$$\sum_{j=1}^{k} X_{ij} \leq 1, \; i = 1, 2, \ldots\ldots\ldots\ldots\ldots\ldots n$$

$$X_{ij} = 0, 1$$

3. SOLUTION TECHNIQUE

The CCZOC model developed above can not be solved in its usual form. However the use of linear decision rule permits the conversion of the chance-constrains into equivalent deterministic constraints. The quantities of distribution function can be easily used to obtain the deterministic equivalents by the above rule.

A chance constraint stating that some variable $\hat{b}$ is no greater than the value of the random variable $\hat{B}$ at least some fraction of the time is written as

$$\text{Prob.}\left[\, \hat{b} \leq \hat{B} \,\right] \geq \alpha \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

$$\text{Prob.}\left[\, \hat{b} \geq \hat{B} \,\right] \leq (1-\alpha)$$

To ensure that Eq.(4) is satisfied, it is sufficient to ensure that the variable $\hat{b}^{(\alpha)}$ of the random variable $\hat{B}$ which is exceed a fraction α of the time:

$$\hat{b} \leq \hat{b}^{(\alpha)}$$

Fig. 1 illustrates the relationships. The fractions and $(1-\alpha)$ are the probabilistic of exceedence and

$$\alpha = 1 - \int_{0}^{\hat{b}^{(\alpha)}} f_{\hat{B}}\,(\hat{b})\, d\hat{b} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

$$= \text{Prob.}\left[\, \hat{b}^{(\alpha)} \leq \hat{B} \,\right]$$

Using the above rule, it is now possible to write the deterministic equivalent of the constraints if the probability density function of the random variable is known.

Let $f_i^{\alpha_i}$ be the time the value that the random variable, R_i has greater than or equal to 100 α_i percent of time, i.e.

$$F_{Rp}\,(f_p^{\alpha_p}) = \alpha_p \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

or $\quad f_p^{\alpha_p} = F_{Rp}^{-1}\,(\alpha_p)$

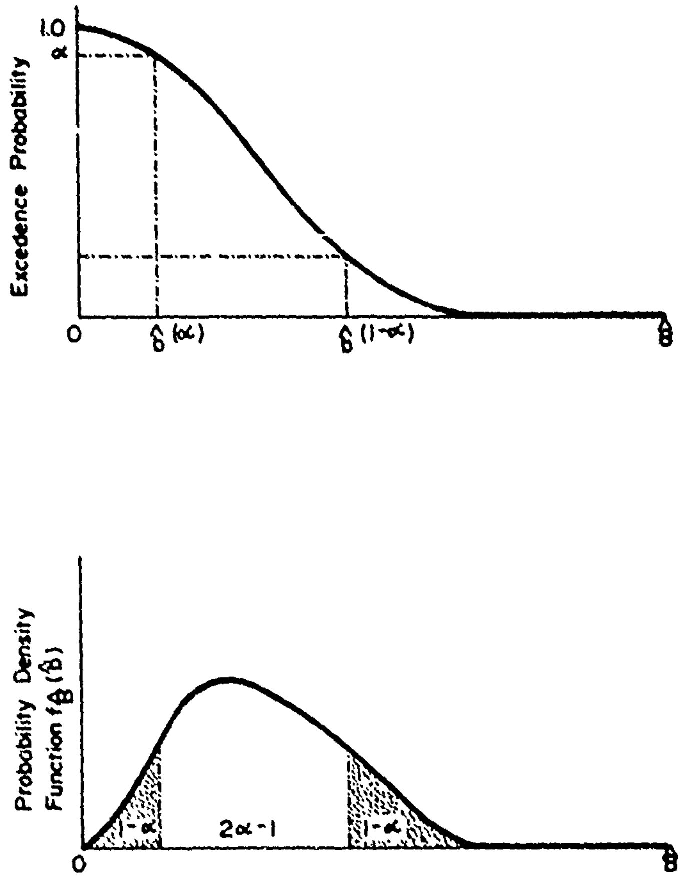


Fig. 1 - Probability Distribution of the Random Variable $\hat{B}$ and Particular Values $\hat{b}^{(\alpha)}$ and $\hat{b}^{(1-\alpha)}$ whose Probabilities of Exceedence are α and (1- α), Respectively.

3.1. The Final Model

The final form of the model can be written as

$$\text{Maximize} \quad \sum_{i=1}^{n} \sum_{j=1}^{k} C_{ij} \; X_{ij}$$

Subject to

$$F_{Rp} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} a_{ijp} \; X_{ij} \right] \geq \alpha_p, \; p = 1, 2, \ldots\ldots m \quad (8)$$

$$\sum_{j=1}^{k} X_{ij} \leq 1 \; , \quad i = 1, 2, \ldots\ldots\ldots\ldots n$$

$$X_{ij} = 0, 1$$

The above model could further be written as

$$\text{Maximize} \quad \sum_{i=1}^{n} \sum_{j=1}^{k} C_{ij} \; X_{ij}$$

Subject to

$$\sum_{i=1}^{n} \sum_{j=1}^{k} a_{ijp} \; X_{ij} \leq f_p^{\alpha_p}, \quad p = 1, 2, \ldots\ldots\ldots\ldots m$$

$$\sum_{j=1}^{k} X_{ij} \leq 1, \quad i = 1, 2, \ldots\ldots\ldots\ldots\ldots\ldots n$$

$$X_{ij} = 0, 1$$

The above problem can hence be solved by zero-one integer algorithm which is commonly available.

4. CONCLUSIONS

In resource constrained capital budgeting problems where the resources are probabilistic variables, it is more logical to formulate probabilistic optimization model like the chance-constrained zero-one programming model developed in this paper which can be used to choose an optimal combination from a large number of projects with many policies. The model can then be converted into a form which is easily solvable by the available algorithm. Such a model also

considers a degree of certainty of meeting the constraints and could provide a tradeoff between the profit and the degree of certainty [3] .

5. REFERENCES

[1] Ahmed, N.U. and D.T. Phillips, A ZERO-ONE INTEGER PROGRAMMING ALGORITHM FOR RESOURCE CONSTRAINED CAPITAL BUDGETING, Proceeidngs of the International Conference on Systems Modelling in Developing Countries, Asian Institute of Technology, Bangkok, pp. 609-626, 1978

[2] Charnes, A. and W.W. Cooper, DETERMINISTIC EQUIVALENTS FOR OPTIMIZING AND SATISFICING UNDER CHANCE CONSTRAINTS, Operations Research, 11, pp. 112-139, 1963

[3] Lohani, B.N. and Thanh, N.C.(1978), STOCHASTIC PROGRAMMING MODEL FOR WATER QUALITY MANAGEMENT MODEL FOR WATER QUALITY MANAGEMENT IN A RIVER, Journal of Water Pollution Control Federation, Vol 50. No. 9, pp. 2175-2182, 1978

[4] Lohani, B.N., A CHANCE CONSTRIANED APPROACH TO REGIONAL WASTEWATER MANAGEMENT SYSTEMS, Proceedings IFAC Symposium on Environmental Systems Planning, Design and Control, Japan, 1977

[5] Lohani, B.N. and Adulbhan, P., GOAL PROGRAMMING APPLIED TO REGIONAL WATER QUALITY MANAGEMENT, Proceedings IFAC Symposium on Environmental Planning, Design and Control, Japan, 1977

[6] Lohani, B.N. and Thanh, N.C., WATER QUALITY MANAGEMENT IN THE HSINTIEN RIVER IN TAIWAN, Journal of American Water Resources Association, Vol. 14, No. 3, 1978

LOAN PORTFOLIO ANALYSIS UNDER PROBABILITY CRITERIUM OF A TYPICAL AGRICULTURAL CREDIT INSTITUTION IN THE PHILIPPINES: A CASE STUDY

Victor A. Tan and Henry J. E. Mandig

Technical Board for Agricultural Credit
Central Bank of the Philippines
Manila, Republic of the Philippines

ABSTRACT. The study is basically an application of the theory of game in respect of agricultural loan activities of a typical agricultural credit institution with focus on Masagana 99. The study also includes regional and loan repayment rates dimensions as prime determinants in an effort to serve the credit needs of small farmers and at the same time maintaining the financial viability of the credit institution.

Parametric sensitivity analysis shall be resorted to with loan volume and repayment levels as the main objects of discourse. Data are derived from the amount of loans granted and collection performance for the seven (7) phases of the Masagana 99 supervised credit program.

Credit allocation ratios decisions provide the conclusory portion of the study. These decisions are made following which appropriate actions should be taken given credit allocation alternatives. Suggestions are in order focusing on the magnitude of loan budget and loan collection efforts which shall be sustained if the credit institution is to continue to exist financially and remain as a contributory factor in small farmer development.

1. INTRODUCTION

The current government thrust in agricultural production in the Philippines has found agriculture credit banks harnessing much of their resources for small farmer supervised[1] credit programs. Nonetheless, several problems have also been noted undermining any gains made in production efforts to the effect that a second look into the loan portfolio mix of lending institutions is necessary at this point in time.

This paper is thus an attempt to investigate into the lending operations of a typical[2] agricultural credit institution especially in respect of the supervised rice production credit program instituted by the government in the early 70's at the time when the attainment of self-sufficiency in rice was the government's prime consideration in most agricultural plans and programs.

To date, loans granted to rice production since the initiation of the program have reached nearly ₱3.5 billion constituting more than 60 per cent to total loans under the various supervised credit programs. The economy is now a rice-surplus and thus a rice-exporting country.

The rice production credit program still obtains today but has assumed a new dimension as it is primarily geared towards a multi-pronged purpose of developing the small farmers into self-financing economic units, monetizing the rural countryside and enhancing savings mobilization and capital accumulation. The first integrated agricultural credit plan (IACP CY 1977-1982) underscores this paramount goal but thus far, much has yet to be desired in this regard.

---

1. By supervised is meant the following features are present: a) application of the package of technology (POT) prescribed by the government; b) farm plan and budget; and c) technical supervision by technicians.

2. By typical is meant an institution whose extent of loanable funds exposure to agricultural lending is more than 80 per cent and whose authorized capitalization is no more than ₱250 million.

Most credit programs in agriculture are supported financially by at least three types of financial institutions: the unit rural banks which are privately owned numbering 830 at the latest count scattered all over the country; the Philippine National Bank, a state-owned and largest commercial bank; and the Agricultural Credit Administration--a non-bank lending institution, likewise state-owned. The unit rural banks and the Agricultural Credit Administration fall under the typical institution category and this paper envisions to delve on the lending activities of these representative credit conduits.

2. THE PROBLEM

The problem of loan arrears by farmers has, however, strained the financial schemes of lending institutions. Latest statistics reveal that outstanding loan defaults have reached over ₱870 million in Masagana 99 credit program alone. This constitutes a high 25 per cent of total loans granted.

Reasons for this relatively high defaulting rate are varied and many but most have been found to be caused by low farm output due to floods and typhoons, loan inadequacy and farmer's attitude towards credit, not necessarily in this order. These are compounded by lack of other sources of income and mismanagement of funds by farmers.

This situation deserves a second look because these farmers' problems are translated into lending institutions' problems as well and finally to the Central Bank of the Philippines. Why this is so springs from the fact that agricultural financial institutions, once plagued by borrowers' non-repayment would seek redress with the Central Bank through a plan of payment for their own arrears--this, in addition to the Bank's other direct financial assistance such as rediscounts and loan advances in the form of special time deposits (STDs) both at low preferential interest rates.

Nonetheless, this past-due problem has reached considerable proportions to the effect that the government has

assumed the role of a fund source agent as well as a subsidizer-lender. The problem therefore has to be worked out within the lending portfolio policy of agricultural credit institutions.

3. THE NEED FOR LOAN PORTFOLIO ANALYSIS

In view of the above considerations, the need for loan portfolio analysis is an expedient and rational preoccupation among agricultural credit institutions if the use of scarce funds as credit is to be effectively utilized. This crucial need, thus, arises from the following considerations:

3.1. Farmers' Profile

Most of the loan beneficiaries in agriculture are small tillers, tenant-lessees, amortizing owners and other small farmers. On this score, average landholding per farmer is no more than 1.7 hectares especially for rice. This observation implies that the bankability level of these prospective loan recipients does not seem to find favorable treatment by lending banks. The farmers are also found to be beset by cash flow problems especially in the case of agrarian reform beneficiaries where payments of irrigation fees, land amortization, barrio guarantee fund and other contributions under the agrarian reform and supervised credit programs have aggravated farmer's repayment capacity.

3.2. High Risks Nature of Agricultural Activities

Rice and other agricultural production activities are predominantly high-risks ventures. On the average, the country experiences some eighteen typhoon visits a year which mostly affect rice growing areas. Floods and pests also occur regularly. This is one obvious reason for most lending banks to shun away from servicing the credit needs of small farmers.

3.3. Presidential Decree 717

Presidential Decree 717, made into law in May 1975 requires all banks to allocate 25 per cent of total loanable funds generated for agricultural lending. The decree fur-

ther stipulates that 10 per cent of these funds should be earmarked for agrarian reform beneficiary lending and the remaining 15 per cent for general agriculture. Many banks have undercomplied with this requirement mostly due to lack of agricultural lending orientation and fear of low repayment rates.

3.4. Farmers' Attitude Towards Credit

Also dubbed as one of the more important reasons for poor collection performance of most agricultural banks is the farmers' attitude towards credit. This is referred to as dole-out mentality by farmers which sometimes extends to their preference of paying first their loans to the traditional money lenders and loans from banks last.

4. THE CONCEPT

In view of the foregoing, it becomes obvious that a typical agricultural credit institution should at least devise a scheme whereby its funds are finding their way back.

The concept proposed here is a variation of the Bayesian theory as a tool in loan portfolio analysis and as is specifically applied in the Masagana 99 Credit program. The theoretical frame is tabulated below:

Table 1. Credit Allocation Contingency Table

State of Nature= Fixed Loan Portfolio	Occurence Probability=Repayment Rate	Possible Options	Possible Results
B	$R_1 R_2 \ldots\ldots\ldots R_{12}$	$C_1 C_2 C_3 . C_{12}$	$R_n = \sum_{i=1}^{12} \Gamma_r R_i$; $r = 1, 2, \ldots 12$
	R_i	$C_i > R_i$	$C_i = \sum_{j=1}^{12} C_{kj} \Gamma_{Cr}$; $r = 1, 2, \ldots 12$
	(R_{kj}, F_{Rr})	(C_{kj}, F_{Cr}) $C_{kj} = R_{Kj}$	

The above configuration states that with a given budget constraint B, Bank X will have to allocate this among the regions with the respective regional repayment rate as set parameters. This means that with a given set of regional allocation ratios, the scheme assumes a set of repayment rates such that the national average repayment rate R_n $(=\sum F_r R_i)$ is equal to the sum of the regional repayment rates R_i weighted by credit ratios F_r.

The R_i's are further broken down into the classwise repayment rates R_{kj} corresponding to percentage credit allocation F_{Rr}.

Column (3) suggests what actions to take given occurrence probabilities in column (2). This means that Bank X is able to tinker with these probabilities inductively, i.e., to tinker from the observed classwise repayment rates to attain regional repayment rates through altering credit ratios. It can also be done deductively by setting a target collection rate from which the regional rates can be tinkered and ultimately the class-wise components, again through credit ratios modifications. Column (4) will yield possible results due to deductive or inductive operations on Column (3). This forms the meat of the analysis as it treats classwise allocation derived from each classwise repayment rates in order to attain C_i now transformed into a repayment strategy.

The purpose of showing the relationship between what is observed (R_i) and what should be (C_i) is the fact that both may be compared both at the regional level of credit allocation disaggregation and at the tier-wise level of analyses. The (C_{kj}, F_{Cr}) tandems may then be analyzed or sensitized to result in C_i from which the (R_{kj}, F_{Rr}) pairs may also be compared. The R_is are assumed here to be probabilities observed during Phases I through VII. If these probabilities in fact hold, it means that the F_{Rr}'s are tinkerable to yield a better C_i. The C_{kj}-R_{kj} equality is a subsequent result of the repayment probabilities as fixed occurences. And the C_i-R_i inequality is an offshoot of R_{Rr}-F_{Cr} inequality.

The main ingredient of this study is therefore the analysis of changing class-wise credit allocations. This is nothing new and abstracts from highly sophisticated financial analysis which is sometimes not only very unwieldy but also unworkable.

All in all, the scheme is a modified version of the Bayesian principle, the modifications occuring in the class-wise level which ultimately influence the C_i levels and therefore the R_n level. Another modification is the fact that the Bayesian theorem primarily proceeds from purely probabilistic information at all levels while the suggested scheme proceeds from a desired set of results as opposed to the former's where results are derived. Deriving results can be done but the application of the scheme here primarily takes off in a deductive manner for reasons argued previously.

5. GENERAL TRENDS IN PHASEWISE COLLECTION RATES

From Table 2A through Table 2F, it can be observed that phase-wise, the concentration of repayment rates varies among the number of farmers, among amount of loans released and per capita amount of loans. This indicates the rather unsatisfactory performance of Bank X in its Masagana 99 program overall, as of June 30, 1976.

5.1. Phasewise Collection Rates

The same tables suggest that for Phase 1, the bulk of loans and farmers went to those whose propensity to pay was relatively higher. At the first glance, it is noted that 42.7 per cent of farmer-beneficiaries accounting for 36.6 per cent of total loans averaged 90-100 per cent repayment rates. This was followed by a sizeable number of farmers and therefore an equally sizeable amount of loans characterized by repayment rates ranging from 60 per cent to 90 per cent. Down below the number of farmers comprised merely 2.8 per cent and amount of loans 1.7 per cent of total. In the same fashion, it is observed that those who had very low repayment rates had likewise relatively lower loans per borrower value. Exceptions were the 90-100 per cent rates

Table 2A

BANK X's PROGRESS REPORT ON PHASE I OF THE "MASAGANA 99"
Rice Production Program as of June 30, 1976

Repayment Rates (In per cent)	No. of Farmers	% Farmers	Loans Released (in ₱000)	% Loans	₱ Loans/Borrower
0 - 40	200	1.1	157.7	0.6	788.50
40 - 50	295	1.4	234.0	1.1	793.20
50 - 60	0	0	0	0	0
60 - 70	4,486	22.0	3,925.7	18.6	875.10
70 - 80	2,482	12.2	2,758.3	13.1	1,111.30
80 - 90	4,197	20.6	6,318.1	30.0	1,505.40
90 - 100	8,686	42.7	7,730.6	36.6	890.00
Ave. repayment rate = 83.1					
TOTAL	20,346	100.0	21,124.4	100.0	1,010.60

Source: Bank X.

Table 2B

BANK X's PROGRESS REPORT ON PHASE II OF THE MASAGANA 99
Rice Production Program as of June 30, 1976

Repayment Rates (In per cent)	No. of Farmers	% Farmers	Loans Released (in ₱000)	% Loans	₱ Loans/Borrower
0 - 40	1,335	12.5	1,226.6	10.6	918.80
40 - 50	2,728	25.6	3,968.4	34.3	1,454.70
50 - 60	322	3.0	498.9	4.3	1,549.40
60 - 70	2,814	26.4	2,819.6	24.4	1,002.00
70 - 80	1,769	16.6	1,317.2	11.4	744.60
80 - 90	336	3.2	379.1	3.3	1,128.30
90 - 100	1,353	12.7	1,371.8	11.7	1,013.90
Ave. repayment rate = 61.0					
TOTAL	10,657	100.0	11,518.6	100.0	1,072.30

Source: Bank X.

Source: Bank X.

Table 2C

BANK X's PROGRESS REPORT ON PHASE III OF THE MASAGANA 99

Rice Production Program as of June 30, 1976

Repayment Rates (In per cent)	No. of Farmers	% Farmers	Loans Released (in ₱000)	% Loans	₱ Loans/Borrower
0 - 40	6,823	33.8	12,069.7	39.4	1,769.00
40 - 50	6,131	30.3	9,498.0	31.0	1,549.20
50 - 60	1,291	6.4	1,517.2	5.0	1,175.20
60 - 70	3,337	16.5	4,530.5	14.8	1,357.70
70 - 80	59	0.3	34.9	0.1	591.50
80 - 90	2,018	10.1	2,178.5	7.0	1,079.50
90 - 100	522	2.6	816.4	2.7	1,564.00
Ave. repayment rate = 45.7					
TOTAL	20,181	100.0	30.645.5	100.0	1,503.50

Source: Bank X.

Table 2D

BANK X's PROGRESS REPORT ON PHASE IV OF THE MASAGANA 99

Rice Production Program as of June 30, 1976

Repayment Rate (In per cent)	No. of Farmers	% Farmers	Loans Released (in ₱000)	% Loans	₱ Loans/Borrower
0 - 40	2,253	26.5	3,786.2	27.7	1,680.50
40 - 50	1,522	17.9	2,373.9	17.4	1,559.70
50 - 60	1,594	18.7	3,486.6	25.4	2,187.30
60 - 70	655	7.7	777.5	5.7	1,187.00
70 - 80	1,286	15.1	2,114.1	15.5	1,643.90
80 - 90	687	8.1	819.9	6.0	1,193.50
90 - 100	511	6.0	313.8	2.3	614.10
Ave. repayment rate = 48.4					
TOTAL	8,508	100.0	13,672.0	100.0	1,620.10

Source: Bank X.

Table 2E

BANK X's PROGRESS REPORT ON PHASE V OF THE MASAGANA 99
Rice Production Program as of June 30, 1976

Repayment Rates (In per cent)	No. of Farmers	% Farmers	Loans Released (in ₱000)	% Loans	₱ Loans/Borrower
0 - 40	4,663	47.0	9,392.6	51.1	2,014.30
40 - 50	2,010	20.2	4,569.2	24.9	2,273.20
50 - 60	1,303	13.1	2,380.5	13.0	1,826.90
60 - 70	859	8.7	1,271.4	6.9	1,480.10
70 - 80	69	0.7	89.8	0.5	1,301.50
80 - 90	37	0.4	21.6	0.1	583.80
90 - 100	140	1.4	265.3	1.4	1,895.00
Not Reported	847	8.5	391.1	2.1	461.70
Ave. repayment rate = 36.8					
TOTAL	9,928	100.0	18,381.5	100.0	1,852.20

Table 2F

BANK X's PROGRESS REPORT ON PHASE VI OF THE MASAGANA 99 RICE PRODUCTION PROGRAM

As of June 30, 1976

Repayment Rates (In per cent)	No. of Farmers	% Farmers	Loans Released	% Loans	Loans/Borrower
0 - 40	1,069	30.3	2,373.3	37.6	2,220.10
40 - 50	799	22.7	1,188.5	18.9	1,487.60
50 - 60	863	24.4	1,280.6	20.3	1,483.80
60 - 70	128	3.6	145.5	2.3	1,136.70
70 - 80	164	4.6	450.7	7.1	2,748.20
80 - 90	266	7.5	384.9	6.1	1,447.00
90 - 100	244	6.9	487.2	7.7	1,996.70
Ave. repayment rate = 44 9					
TOTAL	3,533	100.0	6,310.7	100.0	1,923.90

Source: Bank X.

Table 2G

BANK X's PROGRESS REPORT ON PHASE VII
OF THE MASAGANA 99 RICE PRODUCTION PROGRAM

Repayment Rates[a/]	No. of Farmers	Loans Released	₱ Loans/ Borrower
-	1,547	3,011.0	2,200.00

Source: Bank X

group who had a lower ₱890.00 per caput against the 80-90 and the 70-90 at ₱1,505.40 and ₱1,111.30, respectively. On the whole, this is a surprise since barring aside for a while, the 90-100 class which bastardized the trend, rational wisdom would refer to lower repayment rates for higher loans per caput. Nevertheless, the other end of the line appears to counter this argument with an observation that higher loans mean enough leeway for meeting production and non-production expenses and therefore are more viable. Unfortunately, this is not so because higher loans per borrower is offset by larger acreage to till resulting therefore in a lower loan amount per hectare per farmer. The average loan amount per borrower was ₱1,010.60 and average repayment rate, 83.1 per cent. Six months later as at the end of December 1976, repayment rate went up to 88.1 per cent.

In Phase II, farmers under the 40-50 and 60-70 categories prevailed in terms of repayment rates. resulting in a lower average collection rate at 61 per cent (as of June 30, 1976) in relation to Phase I. This was up to 67.5 per cent by December 31, 1976.

Another note is the fact that the number of farmers was reduced by at least 50 per cent from 20,346 in Phase I to 10,657 in Phase II. The amount of loans released followed suit and dropped to ₱11.582 M from a level of ₱21.1 M in Phase I. Nonetheless, Phase II's average loans per borrower registered ₱1.072.30.

In Phase III, the lower average repayment rate of 45.7 per cent realized as of June 30, 1976 was the result of the majority of farmers having repayment rates of 0-40 per cent by 33.8 per cent of farmers and 40-50 per cent by 30.3 per cent. Loans released comprised 30.4 per cent and 31.0 per cent, respectively, for the two repayment rates groups. Average amount of loans per borrower went up to ₱1,503.50 which was nearly 50 per cent higher than Phase II.

The fourth lowest repayment rate was recorded in Phase IV on account of the fact that 26.5 per cent of farmers granted 27.7 per cent of total loans notched repayment rates of 0-40 per cent, 17.9 per cent of farmers with 17.4 per cent loans had 40-50 per cent repayments and 18.7 per cent of farmers and 25.4 per cent of loans had collection rates of between 50 per cent and 60 per cent. Average repayment rate

of the Phase was 48.4 per cent. Level of loans per caput was a high ₱1,620.10 due to further decline in the number of farmers served to only 8,508 and increase in total loans granted of ₱13.7 M. As at the end of December 1976, repayment rate was 63.3 per cent.

Phase V had an average repayment rate of 36.8 per cent (June 30, 1976) and 73.6 per cent as of December 31, 1976. A very high 47 per cent of farmers corresponding to 51.1 per cent of loans granted them recorded 0-40 per cent repayment rates. Another relatively high 20.2 per cent of farmers granted with 24.9 per cent of total loans released posted 40-50 per cent rates of collection. The 50-60 per cent group composed 13.1 per cent of farmers and 13 per cent of loans. In addition, the number of farmers served further dipped to 9,928 against opposite trends in total loans released of ₱18.4 M, thus netting a high amount of loans per capita of ₱1,852.20.

An all time high of ₱1,923.90 per borrower was recorded for Phase VI. This was obviously due to the more rapid decline in the number of farmers served hitting only 3,533 against amounts of loans released of ₱6.3 M also a decline from the previous phase level. Thus, Phase VI had the lowest repayment rate of only 44.9 per cent as of June 30, 1976 and 61.3 per cent as of December 31, 1976, all attributable to the fact that the 0-60 per cent repayment rates groups almost comprised 80 per cent of total farmers and 76.8 per cent of total loans disbursed.

Phase VII as of this writing served a meager 1,547 farmers and lent out ₱3 M in loans, resulting in a per caput level of loans of ₱2,200.00. Latest available statistics show that 100 per cent repayment rate was registered as of December 31, 1976. This conceals the fact, however, that only a small portion of the loans may have matured since Phase VII covered the period May 1976-October 1976, and farmers with matured loans were able to pay. The whole Phase VII program can only be objectively appraised by May 1977 when all loans will have been matured.

Table 2H

BANK X's PROGRESS REPORT ON PHASES I - VII OF THE
MASAGANA 99 RICE PRODUCTION PROGRAM
As of June 30, 1976

Repayment Rates (In per cent)	No. of Farmers	% Farmers	Loans Released (in ₱000)	% Loans	₱ Loans/ Borrower
0 - 40	16,342	22.1	29,006.1	28.5	1,774.90
40 - 50	13,485	17.3	21,832.3	18.7	1,618.30
50 - 60	5,373	7.3	9,163.8	9.0	1,705.50
60 - 70	12,279	16.6	13,470.2	13.2	1,097.00
70 - 80	5,829	7.9	6,765.0	6.6	1,160.60
80 - 90	7,541	10.2	10,345.8	10.2	1,371.90
90 - 100	12,163	16.5	12,985.1	10.8	903.20
Phase VIII	1,547	2.1	3,011.0	3.0	2,200.00
Ave. Repayment Rate = 54.4					
TOTAL	73,859	100.0	101,959.3	100.0	1,493.00

Source: Bank X.

In Table 2H are shown the overall indicators inclusive of all phases. As revealed, the 16,342 farmers served composing a substantial 22.1 per cent of total farmers turned out to be among the 0-40 repayment rates groups, notwithstanding their share to total loans granted of 28.5 per cent or ₱29 M. The overall repayment rate of a low 54.4 per cent was compounded by the 40-50 per cent collection rates group numbering 18,858 farmers or constituting 24.6 per cent of total with corresponding total loans share of 27.7 per cent or ₱31 M. The 60-70 per cent group totalled 5,829 beneficiaries or 7.9 per cent of the total making use of 13.2 per cent of total loans placed at ₱13.5 M. Despite the 90-100 classification group comprising 16.5 per cent (12,163) of total number of farmers availing of 10.8 per cent of total loans, this did not affect very much the overall collection rate.

5.2. Inter-Phase Variability in Collection Rates

In Table 3 is presented the inter-phase comparison of repayment rates and it is obvious that the variability in rates depends on the time duration, notwithstanding of course the extent of collection efforts during this duration.

This paper attempted to explain the behaviour of repayment rates over time from a fairly limited set of data. Regression estimates on the six phases yielded the following results:

$$y = a + b\,x \Rightarrow y = 32.16 + \underset{(9.11)}{7.47x}\ ; \quad R^2 = 71.2 \quad \underset{}{} = 0.05 \qquad (1)$$

$\hat{y}$ - collection rate as of June 30, 1976

$\hat{a}$ - collection rate intercept

$\hat{b}$ - increase in rate per six-month period

x - number of six-month periods

Table 3

REPAYMENT RATE VS. TIME, BY PHASE

Phase	Repayment Rates (June 30 '76)	No. of 6-months	$Y-Y_n$
Phase I (May 1973 - Oct. 1973)	83.1	5.33	+28.7%
Phase II (Nov. 1973 - Apr. 1974)	61.0	4.33	+ 6.6
Phase III (May 1974 - Oct. 1974)	45.7	3.33	- 8.7
Phase IV (Nov. 1974 - Apr. 1975)	48.4	2.33	- 6.0
Phase V (May 1975 - Oct. 1975)	36.8	1.33	-17.6
Phase VI (Nov. 1975 - Apr. 1976)	44.9	0.33	- 9.5
All Phases ($\bar{Y}n$)	54.4		$(Y-\bar{Y}n)=6.5$
Actual Average per phase ($\bar{Y}$) =	53.3		
Estimated Average per phase ($\bar{Y}$)=	53:3		

Source: Bank X.

On the average, the rate of increase in rate of collection was found to be 7.47 per cent per phase or per six-month period or 1.24 per cent per month. The same table suggests that variability is more pronounced for the oldest and relatively newer phases.

Expectedly, if the regression function holds, Phase VII would have reached a repayment rate of 60.8 per cent by May 1977.

Encroaching further into the implications of repayment rates, it would appear in the table that the variability in repayment rates seem to indicate that Bank X may have been serving the same set of farmers and regions despite unfavorable experiences derived from this credit delivery system. Further, it would also appear that due to the rigidity of Bank X's budget, lack of buoyancy in the allocation of loan portfolios and the inadequacy of efficient collection machinery, Bank X finds it difficult to recover losses from other hitherto financiable program areas. These are mere inferences which should be subject to verification but just the same, the Masagana 99 program seems to be affected by these factors. And certainly, future programs following the same trends would likely become a problematic financial feature of the credit system.

5.3. Costs of Non-Repayments

The low repayment rates obviously result in costs to rural banks as well as to farmers. The amount of unpaid loans by farmers could have been used by Bank X for other activities giving reasonable returns. On the farmer's side, the repayment of loans in full would not have burdened them of interest payments since these unrepaid balances are not condoned anyway.

Basing only on the foregone returns that may have accrued to Bank X had the uncollected loans been collected, it will be shown that the amount of foregone income which could have been earned by Bank X had the defaulting farmers

paid their loans in time and in full as shown in Table 4 below:

Table 4. Foregone Income as of December 21, 1976 of Masagana 99 Uncollected Loans

Phase	Phase Duration	Uncollected Loans	Foregone Income[1/]
I	May 1973-Oct. '73	₱ 2.501 M	₱ 0.939 M
II	Nov. 1973-Apr. '74	3.759 M	1.111 M
III	May 1974-Oct. '74	10.149 M	2.231 M
IV	Nov. 1974-Apr. '75	5.013 M	0.747 M
V	May 1975-Oct. '75	4.853 M	0.407 M
VI	Nov. 1975-Apr. '76	2.486 M	0.054 M
TOTAL		₱ 28.761 M	₱ 5.489 M

The ₱5.489 M thus constitutes lost interest payments to Bank X. The same amount constitutes direct costs to the farmers. At this point, the uncollected amount of ₱28.761 M would generate net benefits to the economy if the farmers' income on account of using it is greater than ₱5.489 M. This means that the ₱28.761 M should at least generate ₱34.25 M. Otherwise, if the Bank X, by virtue of employing the ₱28.761 M for the financing activities of other farmers, thus accorded these other farmers additional incomes greater than ₱5;489 M, the excess constitutes net social costs.

---

1. Assumes that rate of interest is 12 per cent compounded annually or 0.949 per cent monthly.

Nonetheless, from Bank X's side, it is a cost and from the farmers' side, a net cost or net benefit depending on whether there is additional income generated.

The analysis can be extended further especially from the farmers' standpoint. And it becomes a complicated and untraceable matter had the unpaid loans been used for consumption instead of production and the estimation procedures would further become more and more complex and beyond recall if consumption expenditures take the form of spending on medicines, children's schooling and the like. By this is meant the determination of future impact of spending on the family which would be reflected in increased family earnings, transition into a higher status or stratum, pride, etc.

5.4. Regional Dimensions

Bank X spreads its Masagana 99 credit resources among all regions even against limited funds when compared with the rural banks and the Philippine National Bank. Most probably, this must be one single reason for its low rate of collection. In short, had it concentrated on a few verifiably high-returning farmers and regions, it might have posted better rates.

The distribution of loans and farmers served, and therefore the per capita loan were fairly uneven from one region to another. And it is observed also that the Mindanao regions were served least both in the number of farmers and in the volume of credit allocated. Differentials in repayments are also noted.

5.4.1. Regional Distribution of Loans vs. Farmers Served

Heavy concentration of loan was accounted for by Central Luzon garnering almost one-third of total loans granted, followed by the Ilocos Region at 13.1 per cent. Almost automatically, the number of farmers served had the same respective proportions. Western Visayas came third in both indicators. Understandably, these three regions have more rice farms than the rest of the regions outside the Mindanao Area. The latter was served the least on account of the volatile political situation beginning 1973.

Table 5A

BANK X's PROGRESS REPORT ON MASAGANA 99

RICE PRODUCTION PROGRAM, BY REGION, PHASES I TO VI

As of June 30, 1976

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region I					
0 - 40	4,756	48.0	8,167.8	61.6	1,720.00
40 - 50	268	2.7	285.8	2.2	1,066.00
50 - 60	-	-	-	-	-
60 - 70	184	1.8	108.2	0.8	588.00
70 - 80	388	3.9	560.6	4.2	1,448.00
80 - 90	-	-	-	-	-
90 - 100	4,331	43.6	4,134.1	31.2	960.0
TOTAL	9,927	100.0	13,256.6	100.0	1,335.0

Table 5A
Page 2

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region II					
0 - 40	2,166	43.4	4,749.5	45.0	2,200.00
40 - 50	853	17.1	1,734.3	16.5	2,033.00
50 - 60	-	-	-	-	-
60 - 70	757	15.2	843.1	8.0	1,114.00
70 - 80	59	1.2	34.9	0.3	600.00
80 - 90	947	19.0	3,095.0	29.4	3,270.00
90 - 100	204	4.1	86.0	0.8	430.00
TOTAL	4,985	100.0	10,542.8	100.0	2,120.00

Table 5A
Page 3

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
<u>Region III</u>					
0 - 40	2,322	11.1	4,008.9	13.0	1,726.00
40 - 50	8,757	41.9	14,244.9	46.1	1,626.00
50 - 60	1,821	8.7	3,876.3	12.5	2,130.00
60 - 70	5,390	25.8	5,369.3	17.4	1,000.00
70 - 80	2,082	10.0	2,312.1	7.5	1,110.00
80 - 90	-	-	-	-	-
90 - 100	489	2.5	1,084.4	3.5	2,218.00
TOTAL	20,861	100.0	30,895.9	100.0	1,481.00

Table 5A
Page 4

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region IV-A					
0 - 40	1,316	22.8	2,742.4	32.1	2,083.00
40 - 50	199	3.4	318.5	3.7	1,600.00
50 - 60	519	8.9	909.7	10.6	1,752.00
60 - 70	1,065	18.4	1,663.7	19.4	1,562.00
70 - 80	479	8.5	787.7	9.2	1,645.00
80 - 90	1,975	34.2	1,809.3	21.2	916.00
90 - 100	225	3.8	326.5	3.8	1,451.00
TOTAL	5,778	100.0	8,557.8	100.0	1,481.00

Table 5A
Page 5

	Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer
Region V	0 - 40	2,866	30.0	3,175.0	34.6	1,110.0
	40 - 50	971	10.0	1,890.3	20.6	1,950.0
	50 - 60	1,084	11.3	780.0	8.5	719.6
	60 - 70	623	6.5	376.2	4.1	603.8
	70 - 80	1,036	10.8	633.8	6.9	611.8
	80 - 90	326	3.4	376.2	4.1	1,154.0
	90 - 100	2,636	28.0	1,944.8	21.2	724.1
	TOTAL	9,592	100.0	9,176.3	100.0	960.0
Region VI	0 - 40	1,865	16.0	2,719.8	24.7	1,458.3
	40 - 50	653	5.6	286.3	2.6	438.4
	50 - 60	245	2.1	176.2	1.6	719.2
	60 - 70	3,962	34.0	3,743.9	34.0	945.0
	70 - 80	617	5.3	671.6	6.2	1,088.5
	80 - 90	2,599	22.3	2,202.3	20.0	847.4
	90 - 100	1,713	14.7	1,211.3	11.0	707.1
	TOTAL	11,654	100.0	11,011.6	100.0	944.9

Table 5A
Page 6

Repayment Rates	No of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region VII					
0 - 40	646	65.0	423.5	83.9	655.00
40 - 50	48	5.0	10.8	2.1	225.00
50 - 60	-	-	-	-	-
60 - 70	101	10.0	37.9	7.5	375.00
70 - 80	-	-	-	-	-
80 - 90	-	-	-	-	-
90 - 100	193	20.0	32.9	6.5	170.00
TOTAL	988	100.0	505.1	100.0	511.00

Table 5A

Page 7

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region VIII					
0 - 40	1,301	44.0	1,213.8	59.6	932.00
40 - 50	76	2.0	40.6	2.0	540.00
50 - 60	733	25.0	424.4	20.8	580.00
60 - 70	522	18.0	291.2	14.3	560.00
70 - 80	-	-	-	-	-
80 - 90	-	-	-	-	-
90 - 100	317	11.0	67.8	3.3	213.00
TOTAL	2,949	100.0	2,037.8	100.0	691.00

Table 5A
Page 8

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region IX					
0 - 40	301	25.0	267.3	24.1	890.00
40 - 50	-	-	-	-	-
50 - 60	326	27.0	275.1	24.8	850.00
60 - 70	140	12.0	107.1	9.6	765.00
70 - 80	337	28.0	363.5	22.7	1,078.00
80 - 90	69	6.0	61.8	5.6	900.00
90 - 100	34	2.0	35.9	3.2	1,055.00
TOTAL	1,207	100.0	1,110.7	100.0	920.00

Table 5A

Page 9

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region X					
0 - 40	35	4.0	60.0	3.8	1,715.00
40 - 50	-	-	-	-	-
50 - 60	189	17.0	249.1	15.9	1,320.00
60 - 70	56	5.0	61.2	3.9	1,100.00
70 - 80	36	3.0	58.7	3.8	1,630.00
80 - 90	220	19.0	295.6	19.0	1,350.00
90 - 100	594	52.0	837.2	53.6	1,410.00
TOTAL	1,130	100.0	1,561.8	100.0	1,382.00

Table 5A
Page 10

Repayment Rates	No. of Farmers Served	% Share	Loans released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region XI					
0 - 40	310	15.3	876.4	24.1	2,827.00
40 - 50	414	20.5	850.5	23.3	2,054.00
50 - 60	-	-	-	-	-
60 - 70	118	5.8	43.5	1.2	368.00
70 - 80	164	8.1	450.7	12.4	2,748.00
80 - 90	437	21.6	471.0	12.9	930.00
90 - 100	581	28.7	951.7	26.1	1,630.00
TOTAL	2,024	100.0	3,643.8	100.0	1,800.00

Table 5A
Page 11

Repayment Rates	No. of Farmers Served	% Share	Loans Released (In ₱000)	% Share	Average Loan Per Farmer (In ₱)
Region XII					
0 - 40	462	10.9	670.4	10.4	1,451.00
40 - 50	527	12.4	656.9	10.2	1,246.00
50 - 60	697	16.4	1,365.4	21.2	1,960.00
60 - 70	228	5.4	413.0	6.4	1,615.00
70 - 80	792	18.6	1,259.4	19.6	1,590.00
80 - 90	1,239	29.0	1,300.9	23.3	1,211.00
90 - 100	312	7.3	572.6	8.9	1,835.00
TOTAL	4,257	100.0	6,439.5	100.0	1,513.00

Also to be noted and too intelligible is the very meager credit going to Central Visayas at a negligible 0.5 per cent and an equally negligible farmers served of 1.4 per cent. The region is a corn producing area and on this score, corn loans have been concentrated.

A total of 73,010 farmers were served for the Phase I-VI duration and a total of ₱101.3 M disbursed. By average loan per borrower, it was the Cagayan Valley at ₱2,120.00 followed by Southern Mindanao at ₱1,800.00. Central Visayas had the lowest at only ₱511.00.

5.4.2. Regional Repayment Rates

The regional repayment rates among regions at a random glance indicate the pervading influence of low rates for some and high rates for others. Substantial degree of repayment rates lopsidedness can be said of most of the regions where tiltations in rates appear to be found only on two extremes - the high and the low. Region I's (Ilocos) 48 per cent of farmers having 0-40 per cent repayment rates coupled with another extreme - the 43.6 per cent of farmers garnering 90-100 per cent repayment have altogether resulted in a 50 per cent repayment rate, this despite the 31.2 per cent and 4.2 per cent of loans in the 90-100 and 70-80 repayment brackets, respectively.

Region II (Cagayan Valley) bespeaks of the preponderance of the 0-40 groups becoming the ultimate influencing weight on the regions overall repayment rate of 46.5 per cent despite the 29.4 per cent share of loans in the 80-90 category made avail of by 19 per cent of farmers.

Largely, on account of the 41.9 per cent of farmers belonging to the 40-50 tier corresponding to the 46.1 per cent of total loans, Region III (Central Luzon) cornered an average repayment rate of 48.7 per cent. The 17.4 per cent of total loans falling within the 60-70 group was offset by the 13 per cent of loans returning only 0 per cent to 40 per cent.

Region IV-A (Southern Tagalog) had a higher all-phase rate of 52.7 per cent on account of the dominating posture of the 21.2 per cent of loans in the 80-90 classification and of the 19.4 per cent share to total loans in the 60-70 category in spite of the 32.1 per cent of loans reflecting a 0-40 per cent repayment rate.

Like the previous fair regions, Region V (Bicol) settled at a low 55 per cent repayment rate. A total of 3,837 farmers comprising 40 per cent and 55.2 per cent of loans granted them likewise settled at a repayment rate range of 0-50 per cent even at the 90-100 rate class contributed 21.2 per cent in loans and 28 per cent in the number of farmers served.

Region VI (Western Visayas) had a higher 59.7 per cent repayment rate overall. This rate is wedged between the 34 per cent of loans in the 60-70 range and 24.7 per cent of loans in the 0-40 with corresponding share of farmers served of 34 per cent and 16 per cent, respectively. Other higher rates registered fell within the 80-90 and 90-100 which altogether were accounted for by 31 per cent of loans and 37 per cent of farmers.

Region VII (Central Visayas), the corn-producing region registered the lowest rate of 32.2 per cent. This is obviously caught by the 83.9 per cent of loans and 65 per cent of farmers falling within the 0-40 range. This is very substantial with respect to the 6.5 per cent of loans and 20 per cent of farmers having repayment rates between 90 per cent and 100 per cent.

Region VIII (Eastern Visayas) had another very low of 34.4 per cent, the second lowest among regions. This is due to the 59.6 per cent of loans of which only 0-40 per cent were returned notwithstanding the 50-60 rates deduced from 20.8 per cent of loans. No loans recovery were registered in the 70-80 and 80-90 groups while in the 90-100, only a very meager 3.3 per cent of loans were collected.

Region IX (Western Mindanao) had 53.3 per cent in repayment, broken down into 0-40 for 24.1 per cent of loans and 25.0 per cent of farmers, 50-60 for 24.8 per cent of loans and 27 per cent of farmers, 70-80 for 22.7 per cent of total loans and 28 per cent of farmers and the rest in the remaining rates slots.

For Region X (Northern Mindanao) had the highest repayment rate so far placed at a high 82 per cent. Approximately 54 per cent of loans granted belonging to the 90-100 set contributed to this high rate together with the 19 per cent of loans in the bracket 80-90 and 15.9 per cent of loans in the 50-60. Nonetheless, this was hardly contributory to the increase in the overall Bank X repayment rate of 54.4 per cent, all-phase.

Region XI (Southern Mindanao) had another low 54 per cent repayment rate more on account of the considerable 47.4 per cent of total loans and 35.8 per cent of farmers found within the 0-50 per cent repayment category despite the 39 per cent of loans and 50.9 per cent of farmers returning between 80 per cent and 100 per cent of loans.

The 32.2 per cent of total loans reimbursing 90-100 per cent for Region XII (Southern Mindanao) was not enough to show the reverse for an increase in repayment rate due mainly to about 41.8 per cent of loans serving 39.7 per cent of farmers having repayment rates from 0 per cent to 60 per cent. Thus, overall performance only reached 58.2 per cent. This is deemed the third highest, though.

Upon the average, the regions did not seem to have repayment rates influencing the loans per caput ratios. This is shown in Table 5-B where repayment rates erratically relate themselves with loans per borrower, nor was there an established pattern between repayment rates and number of farmers served or total amount of loans granted.

6. THE DESIRABLE EMPIRICAL MECHANISM

The foregoing observations redound to a deduction that Bank X Masagana 99 credit program must have fared better

Table 5B

REGIONAL SUMMARY OF TOTAL LOANS RELEASED BY BANK X
FOR PHASES I - VI UNDER THE MASAGANA 99
RICE PRODUCTION PROGRAM, BY REGION
As of June 30, 1976

Region	No. of Farmers Served	Per Cent Farmers Served	Loans Released (₱000)	% Share	Average Loan/ Farmer	Average Repayment Rate
I	9,927	13.6	13,256.0	13.1	1,335.00	50.0
II	4,986	6.8	10,542.0	10.4	2,120.00	46.5
III	20,861	28.6	30,895.9	30.5	1,480.00	48.7
IV-A	5,778	7.9	8,557.8	8.4	1,480.00	52.7
V	9,431	12.9	10,428.2	10.3	1,105.00	55.0
VI	9,472	13.0	12,343.6	12.2	1,303.00	59.7
VII	988	1.4	505.1	0.5	511.00	32.2
VIII	2,949	4.0	2,037.8	2.0	690.00	34.4
IX	1,207	1.7	1,110.7	1.1	920.00	53.3
X	1,130	1.5	1,561.8	1.5	1,382.00	82.0
XI	2,024	2.8	3,643.8	3.6	1,800.00	54.0
XII	4,257	5.8	6,439.5	6.5	1,513.00	58.2
TOTAL	73,010	100.0	101,322.2	100.0	1,390.00	52.2

Source: Bank X.

Table 6

PER CENT REGIONAL DISTRIBUTION OF BANK X
PER LOAN PORTFOLIO ALLOCATION UNDER THE MASAGANA 99
PHASES I TO VI

Region	Per Peso Share (1)	% Share to Total Farmers Served (2)	Unadjusted Total Share (1) + (2)	Adjusted Total Share
I	.08	13.6	21.6	10.8
II	.074	6.8	14.2	7.1
III	.080	28.6	36.6	18.3
IV	.082	7.9	16.1	8.0
V	.088	12.9	21.7	10.8
VI	.095	13.0	22.5	11.3
VII	.051	1.4	6.5	3.3
VIII	.054	4.0	9.4	4.7
IX	.085	1.7	10.2	5.1
X	.131	1.5	14.6	7.3
XI	.086	2.8	11.4	5.7
XII	.094	5.8	15.2	7.6
TOTAL	1.000	100.0	200.0	100.0

Main Source: Bank X.

than what has obtained were it founded on a pre-determined allocation strategy. This paper inspires from this need notwithstanding the equal need to satisfy other requirements not inherent in Bank X's operations.

The most logical action Bank X would do based on the empirical observations, is to tailor out a schema that would yield the maximum repayment rates. And purely on the latter's basis, Bank X would have focused on serving Region X on the average or on other regions which promise returns on the upper level only. And Bank X would have selected these types of regional strategies if given the options to spread its limited resources.

Nonetheless, the Masagana 99 program is not all profits nor farmers' incomes. It is in fact a delicate combination of both plus its effects on the community and economy as a whole. This conditions thus make it imperative for Bank X to streamline to some extent its loan portfolio and as articulated in the beginning, there is therefore some way of meeting a minimum requirement which blends both private and farmers' profitability.

Upon these assumptions, specifically, it is intended he here to issue out a strategy on the basis of the above observations.

6.1. Regional Credit Allocation

In Table 6 it is shown that if Bank X merely bases its credit loan portfolio on repayment rates, then Region X would have shared the biggest at ₱0.131 per peso and Region VII the smallest at ₱0.051 per peso. But then, there is the other financial objective of serving the farmers in the regions. The two taken together results in the adjusted total share for each region.

The above configuration will result in a moderate increase in repayment rate to 53.2, one per cent higher than the 52.2 per cent registered by Bank X. Obviously, this is not good enough mainly because there is no substantial increase in rates. One valid reason forwarded here is the fact that Bank X allocation of credit still takes into

account the number of farmers for each region which as per computation, has a 50 per cent weight in the determination of credit allocation per peso credit available.

All of this proceeds from the assumption that Bank X loan portfolio for Masagana 99 is fixed together with the fixation assumption on the number of farmers to be served per region.

The foregoing suggests that there is no way for Bank X to become viably and financially capable as an effective credit administrator as far as the Masagana 99 program is concerned.

6.2. Intra-Regional Credit Allocation

Based on the given data on regional repayment rates, and admitting at first the 53.2 per cent as the given parameter, then there are many combinations to achieve the overall rate. One way is of course to use the same intra-farmer allocation based on actual repayment rates observed. The other ways would probably refer to more loans to those with high repayment rates and less to lower ones. This analysis is waived until the topic on application of the scheme is tackled which follows this treatment immediately.

7. SCHEME APPLICATION

The application of the scheme is quite simple, straightforward and nothing fancy. It is composed of two levels of disaggregation, one which applies directly on the region and the other on the farmers of said regions.

Obviously, Bank X has to seek for a repayment rate that is feasible enough as to maintain its cash flow and other Masagana 99 operations. This means that from a standpoint of being able to serve itself financially and the farmers, Bank X has to look for a break-even point. This break-even point should be the minimum that must be maintained at any point in time barring aside of course profitability considerations.

For simplicity purposes, let this break-even point be 84 per cent. This is defensible on the ground that Bank X charges 12 per cent nominal rate on its loan and the 4 per cent on other charges, thus resulting in an effective rate of 16 per cent. In other words, the farmer-borrower is charged 16 per cent against his loan proceeds which constitutes income to Bank X. A complete recovery of loans plus interest charges would amount to 116 per cent of amount of loan. Thus, recovery of Bank X's loan would mean looking only for the remaining 84 per cent.

Of course, this is not as simple as it appears especially if other considerations are put ashore including liability-equity posture of Bank X, interest payments on loans, etc. However, it is assumed here that since Bank X is financed budgetarily, its financial structure is governed by the government, the latter working along non-profit motivations.

7.1. Regional Credit Repayment Rates

The imposition of the break-even criterion likewise imposes regional repayment rates much higher than the observed or those hinging on repayment-based consideration asserted here previously.

This mandate is distributed among the regional repayment rates in proportional doses treated as additions in rates to the actual. This is shown in Table 7. These rates will serve as the controlling rates that Bank X must achieve in order to stay alive and granting that the target in the number of farmers is the same as before.

It is also shown in the table that the low-repaying regions need only to forge rates a little higher than 50 per cent while the others have to obtain relatively higher rates than previously.

The process employed here is deduction. This is simpler to operate both mathematically and operationally. Induction is also possible but the main difficulty lies in the fact that if there is a national or regional target it will require plenty of iterations to achieve that target.

Table 7

REGIONAL REPAYMENT RATES UNDER THE
84 PER CENT OVERALL RATE CRITERION

Region	Per Cent Rate
I	80.5
II	74.9
III	78.3
IV-A	84.9
V	88.6
VI	96.0
VII	51.5
VIII	55.5
IX	81.9
X	95.0
XI	86.7
XII	93.6
TOTAL	84.0

Main Source: Table 53

7.2. Inter-Class Regional Credit Allocation

Proceeding from the given in Table 7, it is not possible to prescribe credit allocations in each class given a fixed regional allocation rate.

As contended elsewhere, it is assumed that repayment rates in each class are probability occurrences. This means that in a particular region i, it is expected that M_{i1} farmers would turn out to be in the 1 class. This also means that among all farmers in region i, $\frac{M_{i1}}{M_i}$ is the percentage of farmers having a repayment rate within the 1 range.

In particular, for example, the loaning operations when applied in Region I, expectations are that 1.1 per cent of farmers will have repayment rates in the 0-40 per cent category, 1.4 per cent in the 40-50, 22 per cent in the 60-70, 12.2 per cent in the 70-80, 20.6 per cent in the 80-90 and 42.7 per cent in the 90-100.

The problem therefore is to select the types of farmers to be financed. This, in effect, also means the identification of the types and assigned values of the characteristics of the farmers impinging on such specific factors as financial and non-financial ratios.

8. CONCLUSIONS AND RECOMMENDATIONS

From the exercise, several conclusions and recommendations may be derived:

1. Bank X has to streamline its lending policy on supervised credit programs;
2. Farmers must be educated on credit consciousness and credit management;
3. Bank X should increase its lending portfolio and diversify its lending activities; and
4. Bank X should intensify its campaign on loan collections.

9. REFERENCES

1) Technical Board for Agricultural Credit, Performance of Small Farmer Credit, Central Bank, Manila, 1977.

2) Tan, V. A., Cash Flow Systems Analysis of Agrarian Reform Beneficiaries in the Philippines, Asian Institute of Technology, Bangkok, May 1978.

3) Technical Board for Agricultural Credit, Study on Masagana 99 Drop-outs in Nueva Ecija, Central Bank, Manila, 1977.

10. ACKNOWLEDGEMENTS

Profound acknowledgements are due to the following:

1. Dr. Chita Tanchoco-Subido, Executive Director, Technical Board for Agricultural Credit for allowing us to write on an important topic such as this;

2. Beulah dela Peña, Staff, Technical Board for Agricultural Credit for seeing to it that the mathematical and statistical framework are meaningful and relevant; and

3. Ester A. Gonzales, Staff, Technical Board for Agricultural Credit for typing the paper.

ON A GENERALIZED ORDERING POLICY

SHUNJI OSAKI

Department of Industrial Engineering,
Hiroshima University,
Hiroshima 730, Japan

ABSTRACT. In replacement policy, it is generally assumed that there are an unlimited number of spares available for replacement. If a spare can be delivered by an order with lead time, the so-called "ordering policy" takes place. In this paper we discuss a generalized ordering policy which includes several interesting ordering and replacement policies. Introducing the costs for ordering, shortage and inventory, and the discount factor of continuous type, we derive the total discounted expected cost which should be minimized as a criterion of optimality. In a main theorem, we show that the optimum policy should be reduced to two typical ordering policies. We further show that two typical ordering policies are discussed thoroughly. Finally, we show that the generalized ordering policy includes several interesting ordering and replacement policies as special cases.

1. INTRODUCTION

Barlow and Proschan [3] discussed some interesting replacement policies. In replacement policy it is generally assumed that there are an unlimited number of spares available for replacement. If a spare can be delivered only by an order with lead time, the so-called "ordering policy" takes place. In this paper we discuss a generalized ordering policy.

Let us briefly sketch the bibliography of ordering policies. Many researchers (see, e.g., [4], [5], [9] and [10]) proposed modified age replacement policies based on typical age replacement policies by Barlow and Proschan [3]. On the other hand, Wiggins [12] discussed an ordering policy with exponential lifetime distribution. Further, Allen and D'Esopo [1, 2] proposed another ordering policy in which a failed unit is repaired with some probability. In this paper, we consider an ordering policy which is somewhat different from the earlier contributions, and is a straightforward extension of an age replacement policy. In particular, we finally show that our ordering policy includes several interesting ordering and replacement policies as special cases.

2. MODEL

Consider a generalized ordering policy in which a system works by an operating unit. The operating unit obeys an identical and independent lifetime distribution $F(t)$ $(t \geq 0)$ which is *arbitrary*. Each delivered spare can take over its operation with the same distribution. An order for spare is made at a prespecified time t_0, and the ordered spare can be delivered after a constant lead time L and is in standby up to another prespecified time t_1 $(> t_0 + L)$. Of course, if the operating unit fails before t_0, the order is made immediately. On the other hand, when the order is made regularly at t_0, the operating unit fails before t_1, the spare in standby takes over its operation immediately (if the spare is available), or the spare takes over its operation just after the spare is delivered (if the spare is not available).

Let us introduce the costs: An ordering cost c_1 is suffered for each expedited order before t_0 and an ordering

cost c_2 is suffered for each regular order at t_0, where it is plausible to assume that $c_1 > c_2 > 0$. The shortage cost k_1 (≥ 0) is suffered during the time duration when failure of an operating unit takes place and the spare is not available. Finally, the inventory cost k_2 (≥ 0) is suffered during the inventory period of the delivered spare. The costs c_1 and c_2 are suffered for each replacement and the costs k_1 and k_2 are the linear ones depending on the time duration.

Let us introduce the discount factor α of continuous (exponential) type (see, e.g., Van Horne [11]). That is, a unit of cost is discounted by $\exp(-\alpha t)$ after the time duration t.

Define a one-cycle from a replacement to the next replacement. The discounted expected cost during one-cycle is given by

$$\begin{aligned}\phi_\alpha(t_0, t_1; L) &= c_1\int_0^{t_0} e^{-\alpha(t+L)}dF(t) + c_2 e^{-\alpha(t_0+L)}\,\overline{F}(t_0)\\ &+ k_1[(1 - e^{-\alpha L})\int_0^{t_0} e^{-\alpha t}\, F(t)dt + \int_{t_0}^{t_0+L} e^{-\alpha t}\, F(t)dt]\\ &+ k_2\int_{t_0+L}^{t_1} e^{-\alpha t}\,\overline{F}(t)dt, \qquad (1)\end{aligned}$$

where $\overline{F}(t) \equiv 1 - F(t)$. A unit of cost is discounted by

$$\begin{aligned}\delta_\alpha(t_0, t_1; L) &= \int_0^{t_0} e^{-\alpha(t+L)}dF(t) + \int_{t_0}^{t_0+L} e^{-\alpha(t_0+L)}dF(t)\\ &+ \int_{t_0+L}^{t_1} e^{-\alpha t}dF(t) + e^{-\alpha t_1}\,\overline{F}(t_1), \qquad (2)\end{aligned}$$

after one-cycle. Then, the total discounted expected cost for an infinite time span is given by

$$C_\alpha(t_0, t_1; L) = \phi_\alpha(t_0, t_1; L)/[1 - \delta_\alpha(t_0, t_1; L)], \quad (3)$$

which we should minimize as a criterion of optimality. It is noted that

$$1 - \delta_\alpha(t_0, t_1; L) = 1 - e^{-\alpha L} + \alpha\int_0^{t_0} e^{-\alpha(t+L)} \overline{F}(t)dt + \alpha\int_{t_0+L}^{t_1} e^{-\alpha t} \overline{F}(t)dt. \tag{4}$$

3. OPTIMUM ORDERING POLICY

We seek the optimum ordering policy which minimizes (3). If the ordered spare is delivered at $t_0 + L$, the delivered spare is in standby up to time t_1 ($\geq t_0 + L$). The following main theorem asserts that t_1 should be $t_1 = t_0 + L$ or $t_1 = \infty$.

[Theorem 1] For a fixed t_0, the optimum t_1^* which minimizes (3) is $t_1^* = t_0 + L$ if $M_\alpha(t_0; L) \geq 0$ or $t_1^* = \infty$ if $M_\alpha(t_0; L) \leq 0$, where

$$\begin{aligned} M_\alpha(t_0; L) &= k_2[(1 - e^{-\alpha L}) + \alpha\int_0^{t_0} e^{-\alpha(t+L)} \overline{F}(t)dt] \\ &- k_1\alpha[(1 - e^{-\alpha L})\int_0^{t_0} e^{-\alpha t} F(t)dt + \int_{t_0}^{t_0+L} e^{-\alpha t} F(t)dt] \\ &- c_1\alpha\int_0^{t_0} e^{-\alpha(t+L)} dF(t) - c_2\alpha e^{-\alpha(t_0+L)} \overline{F}(t_0). \end{aligned} \tag{5}$$

[Proof] We have

$$\frac{\partial C_\alpha(t_0, t_1; L)}{\partial t_1} = \frac{e^{-\alpha t_1} \overline{F}(t_1) M_\alpha(t_0; L)}{[1 - \delta_\alpha(t_0, t_1; L)]^2}. \tag{6}$$

Note that $M_\alpha(t_0; L)$ is independent of t_1 and $t_1 \geq t_0 + L$. If $M_\alpha(t_0; L) \geq 0$, $\partial C_\alpha(t_0, t_1; L)/\partial t_1 \geq 0$ implies $t_1^* = t_0 + L$. Otherwise, if $M_\alpha(t_0; L) \leq 0$, $t_1^* = \infty$. Q.E.D.

From Theorem 1, we can conclude $t_1^* = t_0 + L$ or $t_1^* = \infty$ according to $M_\alpha(t_0; L) \geq 0$ or $M_\alpha(t_0; L) \leq 0$, respectively.

Define the following two functions:

$$q_\alpha^{I}(t_0; L) = [(c_1 - c_2)r(t_0) - \alpha c_2 + k_1 R(t_0)]$$
$$\times[1 - \delta_\alpha(t_0, t_0+L; L)] - \alpha\phi_\alpha(t_0, t_0+L; L), \qquad (7)$$

and

$$q_\alpha^{II}(t_0; L) = [(c_1 - c_2)r(t_0) - \alpha c_2 + (k_1 + k_2)R(t_0) - k_2]$$
$$\times[1 - \delta_\alpha(t_0, \infty; L)] - \alpha R(t_0)\phi_\alpha(t_0,\infty; L), \qquad (8)$$

where $r(t) = [dF(t)/dt]/\overline{F}(t)$ and $R(t) = [F(t+L) - F(t)]/\overline{F}(t)$ are called the *failure rates* and have the same monotone properties (see Barlow and Proschan [3], p. 23). The superscripts I and II are referred to as "Policy I" ($t_1^* = t_0 + L$) and "Policy II" ($t_1^* = \infty$), respectively.

The following theorem is based on Policy I. The proof is omitted.

[Theorem 2] Suppose $M_\alpha(t_0; L) \geq 0$.

1. The failure rate $r(t)$ $(R(t))$ is assumed to be continuous and monotonely increasing.

(i) If $q_\alpha^{I}(0; L) < 0$ and $q_\alpha^{II}(\infty; L) > 0$, there exists a finite and unique optimum t_0^* which satisfies the equation $q_\alpha^{I}(t_0; L) = 0$ and the corresponding cost is given by

$$C_\alpha(t_0^*, t_0^*+L; L) = [(c_1 - c_2)r(t_0^*) + k_1 R(t_0^*)]/\alpha - c_2. \qquad (9)$$

(ii) If $q_\alpha^{I}(\infty; L) \leq 0$, then $t_0^* = \infty$.

(iii) If $q_\alpha^{I}(0, L) \geq 0$, then $t_0^* = 0$.

2. The failure rate $r(t)$ $(R(t))$ is assumed to be continuous and non-increasing. Then $t_0^* = 0$ or $t_0^* = \infty$.

The following theorem is based on Policy II and is given without proof.

[Theorem 3] Suppose that $M_\alpha(t_0; L) \leq 0$ and $C_\alpha(t_0, \infty; L) < k_1/\alpha$ for any t_0 (≥ 0).

1. The failure rate $r(t)$ $(R(t))$ is assumed to be continuous and monotonely increasing.

(i) If $q_\alpha^{II}(0; L) < 0$ and $q_\alpha^{II}(\infty; L) > 0$, there exists a finite and unique optimum t_0^* which satisfies the equation $q_\alpha^{II}(t_0; L) = 0$ and the corresponding cost is given by

$$C_\alpha(t_0^*, \infty; L) = [(c_1 - c_2)r(t_0^*) - \alpha c_2 + (k_1 + k_2)R(t_0^*) - k_2]/[\alpha R(t_0^*)]. \tag{10}$$

(ii) If $q_\alpha^{II}(\infty; L) \leq 0$, then $t_0^* = \infty$.

(iii) If $q_\alpha^{II}(0; L) \geq 0$, then $t_0^* = 0$.

2. The failure rate $r(t)$ ($R(t)$) is assumed to be continuous and non-increasing. Then $t_0^* = 0$ or $t_0^* = \infty$.

4. CONCLUDING REMARKS

If we consider the non-discounted problem, we have the expected cost per unit of time in the steady-state by assuming $\lim_{\alpha\to 0} C_\alpha(t_0, t_1; L)/\alpha$. The corresponding results are rewritten in a similar fashion.

In particular, Theorem 2 is rewritten by assuming $\alpha\to 0$:

$$\lim_{\alpha\to 0} q_\alpha^{I}(t_0; L)/\alpha = [(c_1 - c_2)r(t_0) + k_1R(t_0)] \times[L + \int_0^{t_0} \overline{F}(t)dt] - [c_1F(t_0) + c_2\overline{F}(t_0) + k_1\int_{t_0}^{t_0+L} F(t)dt], \tag{11}$$

and

$$\lim_{\alpha\to 0} \alpha C_\alpha(t_0^*, t_0^*+L; L) = (c_1 - c_2)r(t_0^*) + k_1R(t_0^*), \tag{12}$$

which are given by Osaki [8]. Kaio and Osaki [6] modified the similar ordering policy with salvage cost.

Theorem 3 is rewritten by assuming $\alpha \to 0$:

$$\lim_{\alpha \to 0} q_\alpha^{II}(t_0;\ L)/\alpha = [(c_1 - c_2)r(t_0) + (k_1 + k_2)R(t_0) - k_2]$$
$$\times [ET + \int_{t_0}^{t_0+L} F(t)dt] - R(t_0)$$
$$\times [c_1 F(t_0) + c_2 \overline{F}(t_0) + k_1 \int_{t_0}^{t_0+L} F(t)dt + k_2 \int_{t_0+L}^{\infty} \overline{F}(t)dt], \tag{13}$$

and

$$\lim_{\alpha \to 0} \alpha C_\alpha(t_0^*, \infty;\ L) = [(c_1 - c_2)r(t_0^*) + (k_1 + k_2)R(t_0^*) - k_2]/R(t_0^*), \tag{14}$$

where $ET = \int_0^\infty \overline{F}(t)dt$. The results above are given by Nakagawa and Osaki [7].

It is finally noted that a case of $L = 0$ in Policy I is an age replacement policy with discounting by Fox [5] and a case of $L = 0$ and $\alpha \to 0$ in Policy I is an age replacement policy with no discounting by Barlow and Proschan [3].

REFERENCES

[1] Allen, S.G. and D.A. D'Esopo, AN ORDERING POLICY FOR REPAIRABLE STOCK ITEMS, Opns. Res., Vol. 16, No. 3, pp. 669-674, 1968

[2] Allen, S.G. and D.A. D'Esopo, AN ORDERING POLICY FOR STOCK ITEMS WHEN DELIVERY CAN BE EXPEDITED, Opns. Res., Vol. 16, No. 4, pp. 880-883, 1968.

[3] Barlow, R.E. and F. Proschan, MATHEMATICAL THEORY OF RELIABILITY, John Wiley, New York, 1965.

[4] Cleroux, R. and H. Hanscom, AGE REPLACEMENT WITH ADJUSTMENT AND DEPRECIATION COSTS AND INTEREST CHARGES, Technometrics, Vol. 16, No. 2, pp. 235-239, 1974.

[5] Fox, B., AGE REPLACEMENT WITH DISCOUNTING, Opns. Res., Vol. 14, pp. 533-537, 1966.

[6] Kaio, N. and S. Osaki, OPTIMUM PLANNED MAINTENANCE WITH SALVAGE COST, Int. J. Prod. Res., Vol. 16, No. 3, pp. 249-257, 1978.

[7] Nakagawa, T. and S. Osaki, OPTIMUM ORDERING POLICIES WITH LEAD TIME FOR AN OPERATING UNIT, Revue Francaise d'Automatique, Informatique et Recherche Operationnelle: Recherche Operationnelle, to appear.

[8] Osaki, S., AN ORDERING POLICY WITH LEAD TIME, Int. J. System Sci., Vol. 8, No. 10, pp. 1091-1095, 1977.

[9] Osaki, S. and T. Nakagawa, A NOTE ON AGE REPLACEMENT, IEEE Trans. Reliability, Vol. R-24, No. 1, pp. 92-94, 1975.

[10] Osaki, S. and S. Yamada, AGE REPLACEMENT WITH LEADTIME, IEEE Trans. Reliability, Vol. R-25, No. 5, pp.344-345, 1976.

[11] Van Horne, J.C., FINANCIAL MANAGEMENT AND POLICY, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974.

[12] Wiggins, A.D., A MINIMUM COST MODEL OF SPARE PARTS INVENTORY CONTROL, Technometrics, Vol. 9, No. 4, pp.661-665, 1967.

CAPACITY INSTALLATION OF TWO RELATED EQUIPMENT WITH CONVERSION POSSIBILITY

C. O. Fong
Faculty of Economics and Administration
University of Malaya
Kuala Lumpur
Malaysia

Abstract

The problem in which a firm has to meet the demand for the services of two distinct but related equipment over a planning horizon is considered. An equipment of one type can be converted into an equipment of the other type at some costs. Hence demands may be met by direct capacity installation (expansion) or by conversion from the other type of equipment. Capacity installation and conversion costs are assume to be concave reflecting possible economies of scale in these activities.
The objective is to find a policy of capacity installations and conversions between the two types of equipment such that the present value of the total installation and conversion costs is minimized. The problem is formulated and given a network representation. A dynamic programming algorithm similar to that in [2] is then developed which can be used to solve the problem efficiently.

1. Introduction

The problem in which a firm has to meet the demand for two types of services with two types of equipment available, an expensive general-purpose equipment and a cheaper specialized equipment which could provide only one of the services, has been studied (but not solved) in [5]. The objective of the problem is to determine the policy which minimizes the present value of installation (expansion) costs over an infinite horizon. In [2], it is shown that a discrete time version of this problem can be solved efficiently using a dynamic programming algorithm based on a derived recursive relation. An important generalization of this problem of capacity expansion with specialization is the case where a firm has to meet the demand for the services of two distinct but related equipment. An equipment of one type can be converted into an equipment of the other type at some costs. Practical situations in which this problem of capacity installation with conversion possibility is applicable are many and can be found especially in the areas of transportation, manufacturing and communication. An example is the case of a public railway administration which has to meet the demand for passenger and freight services over a planning horizon. A passenger car can be converted into a freight car at some costs and vice versa. The objective of the organization is to determine a policy of capacity installations and conversions such that the discounted installation and conversion costs are minimized over the planning horizon. Another example is the case of a firm which uses two related manufacturing processes. The machinery of one process can be converted into the machinery of another process with some adjustment and rearrangement. Its objective is again to determine a policy of capacity installations and (if necessary) conversions such that the total discounted cost is minimized. In this paper we formulate the capacity installation with conversion possibility problem into a finite discrete time model with concave installation and conversion cost functions reflecting fixed costs and economies of scale in these activities. Although the problem so formulated is quite different from the inventory models of [8], [9], [10] and the one region and the two region capacity expansion model considered in [6] and [2] respectively, we show that an extension

of the approaches use in these references (in particular that of [2]) can be used to solve the problem efficiently.

In the next section we provide a statement and formulation of the problem. A graphic representation of the problem and some properties of the extreme points are given in §3. The derived extreme point properties are utilized in §4 to derive an efficient dynamic programming algorithm to solve the problem. In the last section we show how the problem can be extended to incorporate initial capacities, backlogging (or short-term leasing) of capacities, arbitrary and not necessarily nondecreasing demands, and the more general case in which there are more than two types of equipment with conversion possibilites from one type to another.

2. Problem Formulation

2.1 Statement of Problem

The problem considered in this paper can be stated formally as follows: A firm has to satisfy the demand for the service of each of two types of equipment over a discrete finite time horizon. The demand for each equipment service is a known nondecreasing function of time. It must also be satisfied exactly (i.e. no inventory or backlogging of the services are allowed) at the end of each period by direct capacity installation or by conversion from the other type of equipment. It should be pointed out that for service industries (e.g. the transport industry) to which this model is particularly applicable, this assumption is not restrictive since it is not possible to backlog or hold as inventory the services provided by the equipment [7, pp 329]. The initial capacity of each equipment is assumed zero and capacity may be installed or converted at the beginning of each period. It is assumed that a unit of capacity installed or converted in any time period has a unit service capability until the end of the time horizon. Capacity installation and conversion costs are assumed to be concave, reflecting possible fixed costs and economies of scale in these activities.

The problem is to find a policy of capacity installations and conversions between the two types of equipment such that the present value of the total installation and conversion costs in minimized.

Although the focus of this paper is on the problem so stated, in §5 we discuss some important extensions to the problem (e.g. when there are initial capacities or when the demands are not necessarily nondecreasing) and show how these could be handled as well.

2.2. Mathematical Formulation

We begin the formulation process by defining

$J = \{1, 2\}$, the set of two types of equipment.

$K = \{1, 2, \ldots T\}$, the set of time periods where T is the end of the planning horizon.

For $i \in J$, $t \in K$ define

r_{it} = the known increment in demand for type i equipment during period t. Demand is nondecreasing, so $r_{it} \geq 0$. Further demand for the two types of equipment are assumed to be measured in the same units.

x_{it} = the amount of capacity of type i equipment installed in period t. Since capacity levels can only increase we have $x_{it} \geq 0$.

y_{it} = the amount of capacity of type i equipment converted to type j ($j \neq i$) in period to. Obviously we have $y_{it} \geq 0$.

I_{it} = idle or excess capacity in type i equipment in period t. In this formulation, for brevity, we do not permit backlogging (or short-term leasing) of capacity and hence $I_{it} \geq 0$. In § 5 we show how this assumption can be relaxed.

$C_{it}(x_{it})$ = the cost of installing x_{it} units of type i equipment in time t.

$H_{it}(y_{it})$ = the cost of converting y_{it} units of type i equipment in time t.

The functions $C_{it}(\cdot)$ and $H_{it}(\cdot)$ are assumed to be concave, nonnegative and nondecreasing in the interval $[0, \alpha]$, and are further assumed to be expressed in present value form. We also assumed that the cost of maintaining idle capacity is negligible and can be ignored. For notational convenience we let

$$R_i(s,t) = \sum_{\tau=s}^{t} r_{i\tau} \quad \text{and} \quad R(s,t) = \sum_{i\in J} R_i(s,t)$$

The problem of capacity installation and conversion can now be formulated as problem P:

$$(1)\quad \sum_{i\in J} \sum_{t\in K} C_{it}(x_{it}) + \sum_{i\in J} \sum_{t\in K} H_{it}(y_{it})$$

subject to constraints

$$(2)\quad X_{1t} - y_{1t} + y_{2t} + I_{1t-1} - I_{1t} = r_{1t} \qquad \text{for } t\in K$$

$$(3)\quad X_{2t} + y_{1t} - y_{2t} + I_{2t-1} - I_{2t} = r_{2t} \qquad \text{for } t\in K$$

$$(4)\quad I_{i0} = 0 \qquad \text{for } i\in J$$

$$(5)\quad I_{it} = 0 \qquad \text{for } i\in J$$

$$(6)\quad X_{it}, y_{it}, I_{it} \geq 0 \qquad \text{for } i\in J,\ t\in K$$

Objective function (1) determines the minimum total present value of capacity installation and conversion costs. Constraints (2) and (3) represent the capacity balance equation of type 1 and 2 equipment respectively at each time period. For example, constraint (2) expresses the condition that the change in excess capacity level of type 1 equipment from period t-1 to t (i.e. $I_{1t} - I_{1t-1}$) is equal to the capacity increment X_{1t} plus capacity converted into type 1 equipment from type 2 equipment (i.e. y_{2t}) less the capacity converted into type 2 equipment (i.e. y_{1t}) and demand increment r_{it}.

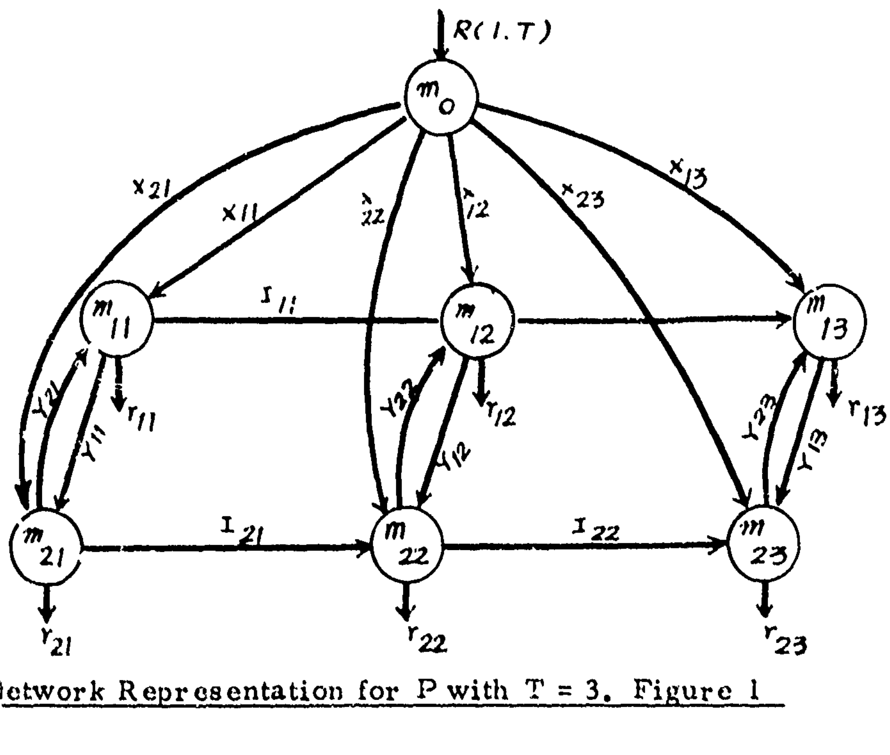
Constraints (4) and (5) reflect the initial and terminal conditions respectively of the capacity states.

Let G represents the constraint set (2) - (6). G is a closed convex set bounded from below and the cost functions are defined such that P has a finite minimum. It is well known [3] that an optimal solution of a concave function over a closed convex set occurs at an extreme point of the constraint set. Since objective function (1) is a concave function and P has a finite minimum, an optimal solution of problem P, therefore, occurs at an extreme point of G. We next derive some properties of extreme points of G which can used to develope an efficient dynamic programming algorithm for solving the problem.

3. Properties of Extreme Point

3.1 Graphic Representation

It can easily be verified that G has the structure of a node-arc incidence matrix which gives rise to the graphic representation as presented in Figure 1.



Network Representation for P with T = 3. Figure 1

Each constraint (2) is represented by flows through a node m_{1t}. Similarly each constraint (3) is represented by a node m_{2t}. The node m_o represents the redundant equation obtained by summing up (2) and (3) over $t \in K$. Following the terminology of [2], we shall call an arc <u>active</u> if the flow along it is positive. An arc that is not active (i.e. have zero flow) is called <u>inactive.</u> An extreme flow (<u>extreme point)</u> of G is defined here as a feasible flow whose active arcs contain no cycle [1], [4] implying that there exists an optimal solution to P whose basic variables (active arcs) do not contain a cycle.

The graph representating G is a single-source multi-sink network. From the results of [10] we can derive the following important properties to an extreme point of P. For completeness we also provides a simple proof for the properties.

<u>Lemma 1.</u> Every extreme point of G satisfies

(7a) $x_{it} \cdot I_{it-1} = 0$ for $i \in j,\ t \in K$

(7b) $x_{it} \cdot y_{it} = 0$ for $i \neq j;\ i,\ j \in J,\ t \in K$

(7c) $y_{jt} \cdot I_{it-1} = 0$ for $i \neq j;\ i,\ j \in J,\ t \in K$

<u>Proof.</u> Consider (7a). $I_{i,t-1} > 0$ implies that there exists a path of active arcs linking node m_{it-1} (via some nodes $m_{j\tau}$, $\tau \in \{1, \ldots, t-1$, $j \in J)$ to the only source node m_o. If $X_{it} > 0$ then we have traced a cycle of active arcs which include arcs (m_{it-1}, m_{it}) and (m_o, m_{it}). This proves that both x_{it} and $I_{i,t-1}$ cannot be positive in an extreme point of G. Properties 7(b) and (7c) can similarly be proved.

The graph G representing problem P is different from the network V representing the problem considered in [2]. One major difference is that the network V of [2] is a multi-source multi-sink network, whereas the graph G of problem P is a single-source multi-sink network. It is this difference which enables the properties of Lemma 1 to be developed for problem P, whereas no properties similar to lemma 1 can be derived for the problem considered in [2].

In spite of the differences, we can characterize an extreme point of G by defining, as in [2], three groups of states at each period t for $t \epsilon K+\{T\}$:

Group 0 : $I_{1t} = I_{2t} = 0$

Group 1 : $I_{1t} > 0, \; I_{2t} = 0$

Group 2 : $I_{1t} = 0, \; I_{2t} > 0$

Group 0 contains only one state. Any state of Group 1 and 2 can be described by a one-dimensional variable. Further the maximum feasible value of $I_{1t}(I_{2t})$ in a state of group 1 (2) at time t is R(t + 1, T). This can be easily verified from Figure 1. Representation of the possible levels of I_{1t} and I_{2t}, when at most only one of them can be positive, by states of groups 0, 1 and 2 provides a useful way of developing a dynamic programming algorithm which is discussed next.

4. A Dynamic Programming Algorithm

4.1 Recursive Relations

States of group 0, 1 and 2, however, do not completely exhaust all possibilities since extreme points of G can also satisfy

(8) $I_{1t} \cdot I_{2t} \neq 0$ for $t \epsilon K - \{T\}$.

Equation (8) is comparable, though not exactly similar, to (16) of [2] and an recursive relationship similar to (17) and (18) of [2] can be developed to solve problem P. For clearity, we shall develope the recursive relationship for solving problem P here, and then point out the differences between the developed relationship and equations (17) and (18) of [2].

In the dynamic programming framework, let the time periods be stages and α be the one-dimensional state variable corresponding to the capacity levels associated with the states of group 0, 1 or 2.

From § 3.1 we know that at period t group 0 has only one value while group 1 and 2 has R(1+1, T)levels each. Thus at each t there are P_t (=2R (t+t, T)+1) possible values for α , with each value of corresponding uniquely to a value of I_{1t} (and I_{2t}).

Let $f_t(\alpha)$ be the cost of optimal capacity installation and converion for periods t+1,... T, with I_{1t} (and I_{2t}) uniquely specified by $\alpha \epsilon \{1, \ldots, P_t\}$.
Let $c_{uv}(\alpha, \beta)$ denote the cost of following an optimal policy over periods u+1 to v given that I_{1u} and I_{2u} (I_{1v} and I_{2v}) are specified by α(β)and $I_{1t} \cdot I_{2t} \neq 0$ for t=u+1,..., v-1. We have:-

(9) $f_T(\alpha=1) = 0$

(10) $f_u(\alpha) = \text{Min}\{c_{uv}(\alpha, \beta) + f_v(\beta) | u < v \leq T, 1 \leq \beta \leq P_v\}$

for u=T-1, T-2,...0 and $\alpha=1, \ldots P_u$

where $P_0 = P_T = 1$.

Recursive relations (9) and (10) represent a one-dimensional state variable dynamic programming relationship and an optimal solution to problem P is given by $f_0(\alpha=1)$. For the procedure to be efficient, the subproblems for deriving $c_{uv}(\alpha, \beta)$ must be soluble trivially. We will now show that to be the case.

4.2 Subproblems

In the calculations in a subproblem to derive $c_{u,v}(\alpha, \beta)$ three cases can be clearly distinguish, with corresponding to the states of each of the three groups (as classified in §3.1) falling under one case.

Case (i) is the case where the α corresponds the state of group 0. Here we have $I_{1u} = I_{2u} = 0$. Since (8) is satisfied for periods u+1 to v-1, by Lemma 1 (properties (7a) and (7c)) we have:

$y_{it} = x_{it} = 0$ for $i \in J$, $t \in \{u+z, \ldots, v\}$.

This implies only y_{iu+1} and x_{iu+1} for $i \varepsilon J$ can be nonzero. Hence we now need to evaluate only the three remaining possibilities where there can be no cycle i. e.

a. $x_{1u+1} = R_1(u+1, v) + I_{1v}$, $x_{2u+1} = R_2(u+1, v) + I_{2v}$

with $y_{1u+1} = y_{2u+1} = 0$

b. $x_{1u+1} = R(u+1, v) + I_{1v} + I_{2v}$, $y_{1u+1} = R_2(u+1, v) + I_{2v}$

with $x_{2u+1} = y_{2u+1} = 0$

c. $x_{2u+1} = R(u+1, v) + I_{1v} + I_{2v}$, $y_{2u+1} = R_1(u+1, v) + I_{1v}$

with $x_{1u+1} = y_{1u+1} = 0.$

Since β specifies I_{1v}(and I_{2v}), in each of the three situations the value of the nonzero variables are specified uniquely and the costs are, threrfore, trivially obtained. The minimum of the three costs is $c_{uv}(\alpha, \beta)$.

<u>In case (ii)</u> α corresponds to a state of group 1. Here we have $I_{1u} > 0$, $I_{2u} = 0$. Again by Lemma 1 we have

$$y_{2t} = x_{1t} = 0 \qquad \text{for } t \varepsilon \{u+1, \ldots v\}$$

$$y_{1t} = x_{2t} = 0 \qquad \text{for } t \varepsilon \{u+2, \ldots, v\}$$

But $x_{2u+1} \cdot y_{1u+1} = 0$ by Lemma 1 (property (7b)).

Thus for extreme flow and feasibility we must have only one of two possibilities

i. e. a. $I_{1u} = R_1(u+1, v) + I_{1v}$ with

$x_{2u+1} = R_2(u+1, v) + I_{2v}$, $y_{1u+1} = 0$

or b. $I_{1u} = R(u+1, v) + I_{1v} + I_{2v}$ with

$y_{1u+1} = R_2(u+1, v) + I_{2v}$, $x_{2u+1} = 0.$

In either case the values of the nonzero variables are uniquely specified. The minimum cost ($c_{uv}(\alpha, \beta)$) can thus be obtained easily. 1/

Case(iii) is the case where α corresponds to a state of group 2. This is similar to case (ii) and need not be elaborated further.

Having shown that $c_{uv}(\alpha, \beta)$ can be trivially derived we are thus justified to say that recursive relations (9) and (10) provide an efficient algorithm for solving problem P.

4.3. Remarks

Although in the main problem recursive relations (9) and (10) are similar to recursive relations (17) and (18) of [2], an important difference exists between the two sets of relations in terms of the solution philosophy used and the computational efforts required for solving the subproblems. As demonstrated in §4.2 for recursive relations (9) and (10) each $c_{uv}(\alpha, \beta)$ can be derived after merely three (if α is in case (i)) or two (if α is in case (ii) or (iii)) arithematic operations. However, for recursive relations (17) and (18) of [2] each $d_{u,v}(\theta, \phi)$ can only be derived after $[v - (u+1)]^2$ arithematic operations (see §4.1 of [2]). Hence, the computational efforts required to solve the subproblems in recursive relations (9) and (10) is trivial and constant, whereas the computational efforts required to solve the subproblems in recursive relations of (17) and (18) of [2] is a square function of v and u.

4.4. Simplications

If the demand increments r_{it} are all even integers (including zero), we can as in [2] restrict the value of the state variable α in the main problem to those associated with only even values (including zero) of I_{1t} and I_{2t}. The reader is referred to [2] for the proof.

Another important simplification arises when the demand for each type of equipment can be assumed to grow linearly i.e. $r_{it} = r_i$ for $i \in J$ $t \in K$. This simplification is presented as the following Lemma.

Lemma 2 For a problem P where the demand grows linearly i.e. $r_{it} = r_i$ for $i \in J$, $t \in K$, at any period t we can in the main problem restrict the value of to those associated with the values of I_{1t} (group 1 states) or I_{2t} (group 2 states) which are partial sums of r_1 and r_2 as follows:

(11a) Group 1: $I_{1t} = ar_1 + br_2$, $\quad I_{2t} = 0$

(11b) Group 2: $I_{2t} = br_1 + ar_2$, $\quad I_{1t} = 0$

where $a \in \{1, 2, \ldots T-t\}$ and $b \in \{0, 1, \ldots T-t\}$.

Proof:

Conditions (11a) and (11b) follow from the concept of exact requirement introduced in [9]. These conditions follow from the rationale that $I_{it} > 0$ with $I_{jt} = 0$ implies that in an extreme flow of G there must exist some subset of nodes $\{m_{l_0 \tau_0}, m_{l_1 \tau_1}, \cdots m_{l_K \tau_K}\}$ where $l_1, \ldots, l_K \in J$, $\tau_1 \ldots \tau_K \in \{t+1, \ldots, T\}$ and $l_0 = i$, $\tau_0 = t+1$ with a corresponding total demand that can be met exactly by idle capacit I_{it}. If such a set does not exist, then it implies that there exists at least one node $m_{l\tau}$, $l \in J$, $\tau \in \{t+1, \cdots, T\}$ whose associated demand is partially met from idle capacity I_{it} and partially from another capacity installation x_{pq}, $p \in J$, $q \in \{1, 2, \ldots, \tau\}$ leading to the existence of a cycle of active arcs in G.

Lemma 2 effectively reduces the values of α in the main problem, at any period t, from 2R (t+1, T) + 1 to 2 (T-t) (T-t+1) + 1. This is an important reduction since it reduces the computational effort from one that is dependent on the demand parameters (i.e. r_{it}) to one that is dependent on the planning horizon (i.e. T), which is typically a small number between 10 to 20 years. It should be noted that a simplification similar to Lemma 2 cannot be derived for the problem considered in [2].

A similar version of conditions (11a) and (11b) can also be derived for problem P in which r_{it} are arbitrary and nonnegative. However the number of distinct values which the partial sums of r_{1t} and r_{2t} for $t \in K$ can assume would be so many that it is, in this case, more efficient to enumerate the values of I_{it} (for states of group i) from 1 to

$R(t+1, T)$ for $t \in K$.

5. Extensions

In this section we consider the various possible extensions to the model P.

5.1 Idle Capacity Maintenance Cost

Maintenance cost of idle capacity in most realistic situations is negligible and has been ignored here. However it can, if necessary, be incorporated into model P by defining the cost function $G_{it}(I_{it})$ to represent the cost of maintaining I_{it} units of idle capacity of type i equipment in period t. Function $G_{it}(.)$ is assumed to be concave, nonnegative and nondecreasing in the interval $[0, \infty)$. Objective (1) would now include the terms $\sum_{i \in J} \sum_{t \in K} G_{it}(I_{it})$. Assuming that the problem has a finite minimum, it can be solved by the approach discussed in § 3 and § 4.

5.2 Backlogging of Equipment Capacity

In a situation where short-term leasing of equipment is available, then it may be realistic to incorporate the possibility of backlogging of each or both type of equipment capacity into the problem P. It need to be pointed out that we assumed it is not possible to have an inventory or backlog of the service provided by the equipment. However, inventory of equipment capacity in the form of idle capacity and backlog of equipment capacity in the form of shortterm leasing of capacity are now permitted. This can be achieved by defining: (12) $I_{it} = Q_{it} - L_{it}$, with $Q_{it}, L_{it} \geq 0$ for

$i \in J$, $t \in K$ with the imposed condition that

$$Q_{it} \cdot L_{it} = 0 \qquad \text{for } i \in J,\ t \in K. \tag{12a}$$

The variables Q_{it} and L_{it} now represent the amount of idle and backlog capacity respectively of type i equipment at period t. In terms of the network representation, this involves the addition of arcs (m_{it+1}, m_{it}) to represent L_{it} with Q_{it} now represented by arcs (m_{it}, m_{it+1}).

We now define $F_{it}(L_{it})$ as the cost of backlogging L_{it} units of type i equipment capacity at period t, where $F_{it}(.)$ is again assumed to be concave, nonnegative and nondecreasing over the interval $[0, \infty)$. Objective function (1) now would include the terms $\sum_{i \in J} \sum_{t \in K} F_{it}(L_{it})$.

Since each node m_{it} now has four arcs representing x_{it}, $y_{jt}(j \neq i)$, Q_{it-1} and L_{it} directed into the node, Lemma 1 can be expanded as:

<u>Lemma 3.</u> Every extreme point of P satisfies

(13a) $x_{it} \cdot Q_{it-1} = 0$ for $i \in J$, $t \in K$

(13b) $x_{it} \cdot L_{it} = 0$ for $i \in J$, $t \in K$

(13c) $x_{it} \cdot y_{jt} = 0$ for $j \neq j$, $j \in J$, $t \in K$

(13d) $y_{jt} \cdot Q_{it-1} = 0$ for $i \neq j$; $j \in J$, $t \in K$

(13e) $y_{jt} \cdot L_{it} = 0$ for $i \neq j$; $j \in J$, $t \in K$

(13f) $L_{it} \cdot Q_{it-1} = 0$ for $i \in J$, $t \in K$.

The proof of this Lemma is the same as the proof for Lemma 1 and needs no further elaboration.

At each period t, the number of states of group 1 and 2 (as classified in §3.1) would now have to be expanded to include negative of I_{it}, i.e.

Group 1: $I_{1t} \neq 0$, $I_{2t} = 0$

Group 2: $I_{1t} = 0$, $I_{2t} \neq 0$.

Now states in each group 1 and 2 can assume a maximum of R(1, T) levels. Hence at each period t, we have to enumerate 2R(1, T) +1 levels. The recursive relations (9) and (10) is still applicable except that now

$$P_t = 2R(1, T) + 1 \quad \text{for } t \in K - \{T\} \text{ with } P_0 = P_T = 1.$$

In using recursive relations (9) and (10) the subproblems for calculating $c_{uv}(\alpha, \beta)$ is now slightly more complex.

Referring to the classification of §4.2, in <u>case (i)</u> we can have four subcases:

(14a) $Q_{1t} > 0,\ Q_{2t} > 0,\ L_{1t} = L_{2t} = 0$

for $t = u+1, \ldots, v-1$

(14b) $Q_{1t} > 0,\ L_{2t} > 0,\ L_{1t} = Q_{2t} = 0$

for $t = u+1, \ldots, v-1$

(14c) $L_{1t} > 0,\ Q_{2t} \quad 0,\ Q_{1t} = L_{2t} = 0$

for $t = u+1, \ldots, v-1$

(14d) $L_{1t} > 0,\ L_{2t} > 0,\ Q_{1t} = Q_{2t} = 0$

for $t = u+1, \ldots, v-1$.

As shown in §4.2 for (14a) we have only three possibilities in which we need to compute the costs. For (14b), it can easily be verified by Lemma 3 (using properties (13d) and (13e)) that $y_{2t} = 0$ for $t = u+2, \ldots, v$ and $y_{1t} = 0$ for $t = u+1, \ldots, v-1$. Hence we have only five possibilities in which the costs need to be computed. The five possibilities are

(a) $x_{1u+1} > 0$ with $y_{1v} > 0$,

(b) $x_{1u+1} > 0$ with $x_{2v} > 0$,

(c) $x_{1u+1} > 0$

(d) $x_{2v} > 0$ with $y_{2u+1} \quad 0$ and

(e) $y_{2u+1} > 0$.

Again in each of the five situations the nonzero variables are uniquely specified and the cost can be derived trivially. Conditions (14c) and (14d) are similar to (14b) respectively. Hence for case (i), we have to derive 16 cost data, the minimum of these being $c_{uv}(\alpha, \beta)$.

In case (ii) we also have the four subcases (14a) to (14d). For (14a) it can be shown that, since now I_{1u} can be negative, we now need to compute the costs for five possibilities.2/ In each of (14b), (14c), (14d), we have eight possibilities.3/ Hence to compute $c_{uv}(\alpha, \beta)$ with α from a group 1 state we have to derive 29 cost data and select the minimum.

Case (iii) is similar to case (ii) and will not be further elaborated. We have now shown that the subproblems can still be solved easily for the backlogging case, justifying the fact that recursive relations (9) and (10) can still be applied efficiently.

(5.3) Initial Capacities

Initial capacity in either or both type of equipment can also be incorporated into model P. In this case the state specified by α at stage 0 would correspond to the capacity status of each type of equipment. Recursive relations (9) and (10) can still be used to solve the problem. However now the network representing G is no longer a single-source network and Lemma 1 need not necessarily hold true. Now the subproblem for calculating $c_{uv}(\alpha, \beta)$ is slightly more complicated. Referring to the classification of § 4.2, case (i) remains unchanged since $I_{1u} = I_{2u} = 0$ implies that Lemma 1 still holds for nodes $m_{i\tau}$, $i \in J$ and $u < \tau \leq V$. In case (ii), we still can have at most one installation for each type of equipment between periods u+1 and v. The cost for the possibility of no expansion between u+1 and v is uniquely specified, since for feasibility y_{1u+1} must be nonzero and, hence, y_{1t} for $t = u+2, \ldots v$ and y_{2t} for $t = u+1, \ldots, v$ must be zero. Consider the possibility in which there is only one expansion. When the expansion occurs at node m_{1T} for any T between u+1 and v or m_{2T} for any τ between u+2 and v the cost is uniquely specified since the value the expansion is known and y_{1u+1} 0 for feasibility.

There are, therefore, (v-u) + (v-u-1) such evaluations. When the expansion occurs at node m_{2u+1} we have 2(v-u) possibilities, each with $y_{1\tau}$ or $y_{2\tau}$ (but not both) nonzero for $\tau \varepsilon \{ u+1, \ldots, v \}$. Hence we have a total of 4 (v-u) - 1 possibilities to consider when there is only one expansion. When there are two expansions between periods u+1 and v, there cannot be any capacity conversion between periods u+1 and v. This implies we now have u-v possibilities with x_{2u+1} 0 and $x_{1\tau}$ 0 for any τ between u+1 and v. Hence in this case (and also case (iii)), we have to consider 5(v-u) possibilities, each with the cost uniquely specified.

5.4 Arbitrary Demand

The case in which the demands are arbitrary (i.e. not necessary nondecreasing) may be realistic in some situations. For example, some public railway administrations have experienced an absolute decline in demand for passenger traffic (cars) due to increasing competition of other modes of passenger transportation, while the demand for freight cars have generally maintained a steady growth. The formulation and network representation of problem P still remains valid when the demands are arbitrary. Now we have $R_i\ (1,\ t) \geq 0$ for $i \varepsilon J$ and $t \varepsilon K$, but r_{it} is unrestricted in sign. In the network each $r_{it} < 0$ would be represented by an inflow of r_{it} units into node m_{it}. As in the case with initial capacities, the network is now a multi-source network and Lemma 1 is no longer valid. Recursive relations (9) and (10) can still be used to solve the problem, but the subproblems would have to be solved in the manner described in § 5.3.

5.5 Multiple Types of Equipment

The model given in § 2 and the network representation can be extended to handle the more general case where there are more than two types of equipment with conversion possibilities from any one type of equipment to another. A decomposition property similar to equation (8) can still be derived. However, now at each period t we have k groups of states where k is the total number of types of equipment available. Each group ℓ is now defined as $I_{\ell t} = 0$ with $I_{st} \geq 0$ for $s \varepsilon \{ 1, 2, \ldots, k \} - \{ \ell \}$.

For notational convenience this classification is slightly different from that given in §3.1. The dimensionality of the state variable (i.e. α) in recursive relation (10) is now k-1. The subproblem of §4.2 would now be relatively more difficult to solve, though there can now still be at most one expansion for each type of equipment.
It can, nevertheless, be solved in a manner similar to that indicated in §4.2. When the demand for each type of equipment can be assumed to grow linearly (i.e. $r_{it} = r_i$ for $i \in \{1, 2, \ldots, k\}$ and $t \in K$) then an equivalent form of (11a) and (11b) can be derived which can be used to substantially reduce the number of k-1 dimension state variables α that need to be considered in the recursive relatives (9) and (10). In fact in this case, at each period t, the number of levels of I_{it} that needs to be enumerated is approximately $(T-t)^k$, which implies that the total number of k-1 dimension state variable is of the order $(T-t)^{k(K-1)}$.
In realistic problems T is typically about 15 years. Hency for the case of constant growth rate in demand for equipment capacity, the procedure may be tractable for k relatively small (e.g. $k \leq 4$).

6. Conclusions

In this paper it is shown that a recursive relationship developed in [2] can be adpted to solve the problem of capacity installation for two types of related equipment with conversion possibility from one type to the other - a problem encountered particularly in the transport industry. In adopting the recursive relationship of [2] to solve this problem, it is further shown that the solution philosophy used to solve the subproblems is quite different from, and more efficient than, that used in [2]. Important simplications and extensions to the basic problem considered are also derived.

FOOTNOTES

1. When I_{1u} does not equal the two values specified, the flow is non-extreme and, hence, in these cases $c_{uv}(\alpha, \beta)$ can be set arbitrarily high.

2. The possibilities are:

   (a) $x_{1u+1} > 0$ with $y_{1u+1} > 0$,

   (b) $x_{1u+1} > 0$ with $x_{2u+1} > 0$, and

   (c) $x_{2u+1} > 0$

   (d) $x_{2u+1} > 0$ with $y_{2u+1} > 0$, and

   (e) $y_{1u+1} > 0$.

3. For (14b) the possibilities are (a) no positive installation or conversion variable between periods u+1 and v.

   (b) $y_{2u+1} > 0$, (c) $y_{1v} > 0$

   (d) $x_{1u+1} > 0$, (e) $x_{1u+1} > 0$, $y_{1v} > 0$,

   (f) $x_{1u+1} > 0$, $x_{2v} > 0$, (g) $x_{2v} > 0$,

   (h) $x_{2v} > 0$, $y_{2u+1} > 0$.

For (14c) and (14d) the possibilities are similar and, therefore, need not be given here.

REFERENCES

1. G. B. Dantzing, Linear Programming and Extensions, (Princeton University Press, Princeton, New Jersey, 1963).

2. C. O. Fong and M. R. Rao, Capacity Expansion with Two Producing Regions and concave Costs, Management Science 22 (1975) 331-339.

3. G. Hadley, Nonlinear and Dynamic Programming, (Addison-Wesley Publishing Company, Reading, Massachusetts, 1964).

4. E. L. Johnson, Networks and Basic Solutions, Operations Research 14 (1966) 619-623.

5. A. J. Kalotay, Capacity Expansion and Specialisation, Management Science 20(1973) 56-64.

6. A. S. Manne and A. F. Vernott Jr., Optimal Plant Size with Arbitrary Increasing Time Paths of Demand in: A. S. Manne (ed.), Investment for Capacity Expansion: Size, Location and Time Phasing (M. I. T. Press, Cambridge, Massachusetts, 1967.)

7. J. K. Sengupta and K. A. Fox., Optimization Techniques in Quantitative Economic Models (North - Holland, Amsterdam, 1969).

8. H. M. Wagner and T. M. Whitin, Dynamic Version of the Economic Lot Size Model, Management Science 4(1958), 89-96.

9. W. I. Zangwill, A Deterministic Multi Period Production Scheduling Model with Backlogging, Management Science 13 (1966) 105-119.

10. W. I. Zangwill, A Backlogging Model and a Multi - Echelon Model of a Dynamic Economic Lot Size Production System - A Network Approach, Management Science 14 (1969) 506-527.

NONRENEWABLE ENERGY CONSUMPTION FORECASTING BY GROWTH CURVES

M.N. SHARIF, S.R. KHAN and M.N. ISLAM

Industrial Engineering and Management Division
Asian Institute of Technology
P.O. Box - 2754, Bangkok, Thailand

ABSTRACT. The purpose of this paper is to assess experimentally how well different growth models work in the same application. The case used for analysis is that of forecasting nonrenewable energy consumption patterns. All existing models of "growth with upper limit" are applied to the same set of data for making alternative forecasts. The characteristics of different growth curves are also studied in general terms as well as in this specific application. Moreover, a new growth model for the case of energy consumption has been developed and evaluated. Results of the analysis are presented in a comparative form, so as to help forecastors to discriminate intelligently among the available tools.

1. INTRODUCTION

One of the most important area of research requiring immediate attention in relation to the study of the future has recently been identified to be on issues of methodology - i.e., improvement in existing methods, attempt to develop new methods or expand applications, etc. [18]. One specific recommendation of the above National Science Foundation Study is to "experimentally asses how well different forecasting methods work in the same application". This paper presents one case study on forecasting by growth curves.

Given the overwhelming impact of energy on present-day societies, the selection of the case of forecasting energy consumption patterns needs no further justification. The factor to be noted is the restriction of the scope of the study to consider only the nonrenewable energy resources - coal, oil, gas and nuclear. Moreover, consumption in this study refers to world total.

Studies on different aspects of energy consumption are abundant in the literature. Even in the area of forecasting alone there are many. Wessels et al [21] studied the state of methods of forecasting demand for energy. Nutty [14] gave some estimates on the possible date of running out of oil supply. Austin and Brewer [1] forecasted world energy consumption by interrelating projected population and energy consumption per capita. Cook [5] referred to Hubbert's "depletion-history curves" for energy forecasting. Forecasts on energy consumption are also incorporated in studies on economic and environmental aspects of energy - e.g., Bailey [2], White [22], Watts and Hrubecky [20], MIT Energy Laboratory [15], etc. The forecasting functions used in all studies mentioned so far, are other than the wellknown growth curves, such as, Pearl curve, Gompertz curve, Logistic curve, Floyd curve, etc. However, Lakhani [11] used Gompertz curve in evaluating an environment-saving process in petrolium refinery industry, and Bossert [4] applied Logistic curve to electric utility forecasting.

The choice of a suitable type of function is an important problem in long-term energy consumption forecasting. The past growth in the use of energy resources in the world has been recorded to be exponential [12]. However, since the nonrenewable energy resources are of finite quantity, it is obvious that the choice of suitable type of function should be limited to "growth curves with upper limit". Therefore, section 2. of this paper deals with the kinds and characteristics of the available growth curves with upper limit. In section 3. these curves are used for forecasting and analysis. The functions selected and the results of projection derived from the regression analysis are checked

for statistical reliability. A new model for forecasting energy consumption patterns is then presented in section 4. Finally, section 5. includes a synthesis of forecasts from diverse methods. It is hoped that the analysis will help forecastors to discriminate among the available growth curves with upper limit.

2. GROWTH CURVES WITH UPPER LIMIT

Since past data shows an exponential trend for energy consumption, in this section we will be considering only exponential growth curves which approach an upper limit asymptotically. In all the mathematical expressions given below, Y is the energy consumption, t is the time and L is the upper limit.

2.1 Simple Modified Exponential [9]

$$Y = L - a\, e^{-bt} \qquad (1)$$

where, a and b are positive constants.

For the simple modified exponential curve the logarithm of the slope when plotted against time gives a straight line sloping down to the right.

2.2 Logistic Curve [4, 9, 21]

$$Y = \frac{L}{1 + e^{a-bt}} \qquad (2)$$

where, a and b are positive constants.

For the logistic curve the logarithm of the ratio of the slope to the square of the moving average, when plotted against time, will give a straight line sloping down to the right. This curve grows through a phase of increasing increasingly to one of increasing decreasingly. It is symmetrical about the inflection point.

2.3 Pearl Curve [13]

$$Y = \frac{L}{1 + ae^{-bt}} \qquad (3)$$

where, a and b are positive constants.

Pearl curve has the same slope characteristics as that of the logistic curve. The curve has initial value of zero at $t = -\infty$ and reaches the limiting value L at $t = +\infty$. Inflection point occurs at $t = (\ln a)/b$ corresponding to $Y = L/2$. The curve is symmetrical about the inflection point. Because of the symmetry, 'a' determines where the

the curve will be located on the time axis and 'b' determines the steepness of the rising portion.

2.4 CUMMULATIVE PROBABILITY CURVE [3]

$$Y = \frac{L}{1 + e^{a+bt}} \tag{4}$$

where, a and b are constants.

This is the cummulative probability curve for normally distributed random variables. The curve has the same characteristics as that of the Pearl curve. It is symmetrical about the inflection point having a range from $-\infty$ to $+\infty$. Inflection point occurs at $t = -(a/b)$, where $Y = L/2$.

2.5 GOMPERTZ CURVE [9, 13]

$$Y = L\, e^{-ae^{-bt}} \tag{5}$$

where, a and b are positive constants.

Gompertz curve has a little different slope characteristics than that of the logistic curve. For the Gompertz curve, the logarithm of the ratio of the slope to the moving average (not the square of the moving average) when plotted against time gives a straight line sloping down to the right. Moreover, unlike the Pearl curve, Gompertz curve is not symmetrical. This curve has a steeper growth than the Pearl curve. The curve ranges from an initial value of zero at $t = -\infty$ to the limiting value of L at $t = +\infty$. The inflection point occurs at $t = (\ln a)/b$, where $Y = L/e$.

2.6 FLOYD CURVE OF UNIVERSAL GROWTH [8, 16]

$$\frac{F}{F - f} + \ln \frac{f}{F - f} = a + bt \tag{6}$$

where, F is the upper limit of the figure of merit represented by f and a and b are constants.

This general curve on technological growth was derived by calculating the probability of improving the figure of merit through applied efforts. The Floyd curve is non-symmetrical. Inflection point occurs at $f = F/3$, where $t = (\ln 1/2 + 3/2 - a)/b$.

It can be observed from the above expressions of different growth curves with upper limit, that the Logistic curve (Eq. 2), the Cummulative Probability curve (Eq. 4) and the Pearl curve (Eq. 3) are essentially the same. Thus, we can reduce the number of available curves to four -- Modified Exponential, Pearl, Gompertz and Floyd. Figure 1 shows the general shape of these growth curves with upper limit. With the exception of the Modified Exponential

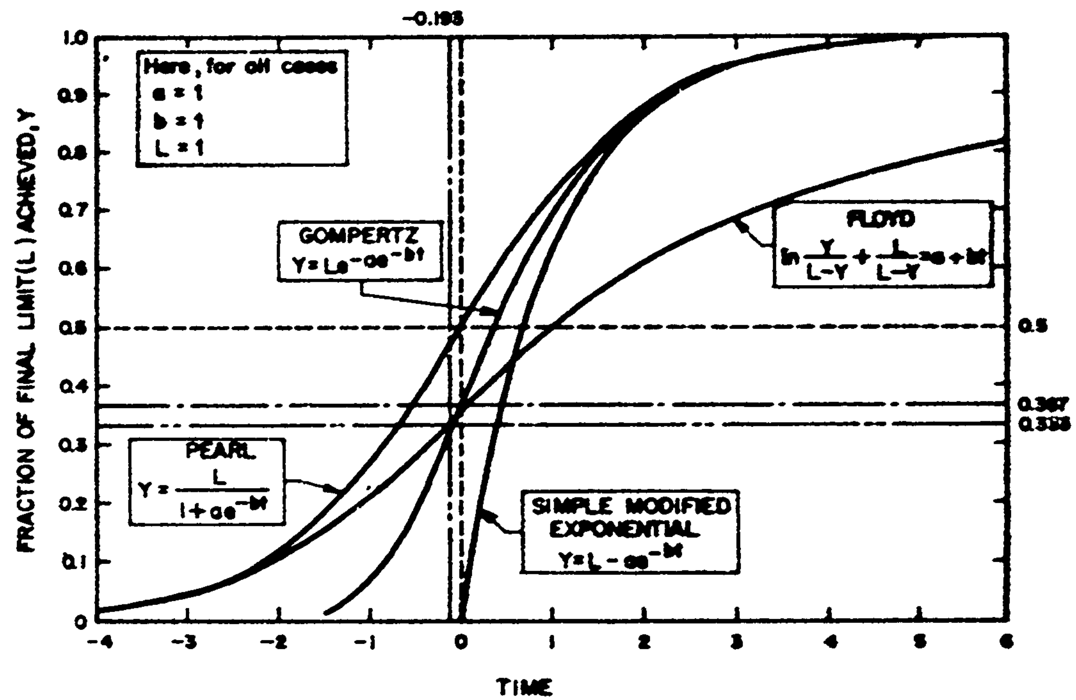


FIG. 1. EXAMPLES OF PEARL, GOMPERTZ, FLOYD AND SIMPLE MODIFIED EXPONENTIAL CURVE.

curve, all others are of the S-shape, i.e., the growth changes from a phase of increasing increasingly to one of increasing decreasingly.

The past observations of exponential growth in world energy consumption conforms to the increasing increasingly phase of the growth curves. However, due to widespread awareness on the depletion of reserves resulting in energy conservation policies, the energy consumption growth is likely to move into the phase of increasing decreasingly. Therefore, the choice of functions for energy consumption forecasting is restricted to the three well-known growth curves - (i) Pearl curve; (ii) Gompertz curve; and (iii) Floyd curve.

3. APPLICATION OF PEARL, GOMPERTZ AND FLOYD CURVES

Historical data regarding world consumption and reserve of primary energy sources are shown in Table 1. Consumption figures are for the period of 1950-1974, and all in terms of million metric tons of coal equivalent. The reserve figures are shown as of 1950 by adjusting for the consumptions since 1950 to the data available on proved and recoverable reserves. Using the data shown in Table 1 and the mathematical expressions for the Pearl (Eq. 3), Gompertz (Eq. 5) and Floyd (Eq. 6) models the curves are fitted by minimizing the sum squares of the ratio of fitted curve and data curve. The resulting curves are shown in Figure 2, and the estimates for the parameters along with the correlation coefficients are given in Table 2.

It can be observed from Figure 2 that the results of long-range projections derived from regression analysis on the basis of the three selected functions are significantly different. Therefore, it is necessary to check for statistical reliability.

The correlation coefficients are observed to be very high for each of the three models (Pearl, Gompertz and Floyd) in each of the four cases (Coal, Oil, Gas and Nuclear). Moreover, for the Pearl and Floyd curves, the logarithm of the ratio of the slope to the square of the moving average gives a straight line sloping down to the right. Also, for the Gompertz curves, the logarithm of the ratio of the slope to the moving average gives a straight line sloping down to the right. Hence, the statistical tests are satisfactory for all the cases. However, since the projections are so different, further analysis is necessary.

Table 1 World Consumption and Reserve of Primary Energy Sources (All Quantities in Million Metric Tons of Coal Equivalent)

CONSUMPTIONS				
Year	Coal	Oil	Natural Gas	Nuclear
1950	1534.188	671.830	244.463	42.200
1951	1643.838	740.449	288.582	45.984
1952	1636.762	787.755	310.790	49.055
1953	1658.734	834.643	329.796	50.699
1954	1666.541	883.815	347.653	53.727
1955	1814.013	991.090	380.040	58.057
1956	1884.270	1073.219	410.553	63.301
1957	1922.601	1117.053	446.554	67.576
1958	1955.044	1187.502	486.624	75.139
1959	2066.790	1267.548	541.995	78.116
1960	2205.717	1358.297	593.637	85.160
1961	2037.224	1440.711	636.580	89.654
1962	2080.896	1556.211	697.483	95.080
1963	2171.490	1680.873	757.023	100.192
1964	2222.450	1815.931	827.579	104.971
1965	2254.516	1954.795	887.298	115.804
1966	2293.168	2137.037	968.942	126.088
1967	2177.273	2269.541	1038.358	129.617
1968	2285.643	2473.732	1136.608	136.242
1969	2353.986	2688.167	1249.412	146.126
1970	2418.260	2935.951	1368.182	154.425
1971	2396.062	3114.285	1468.498	164.076
1972	2406.064	3338.648	1555.187	176.933
1973	2502.904	3597.884	1622.726	185.775
1974	2530.733	3566.777	1668.207	205.098
RESERVES (As of 1950)				
	7888.159	725.483	748.262	4842.599

Sources: (i) World Energy Supplies 1950-1974; U.N. Statistical Office, U.N. Publication, New York, U.S.A. (1975).

(ii) Energy Index 1975 (P-48); Proved and Recoverable Reserve; Environmental Information Centre Inc., New York, U.S.A.

Table 2 Parameters and Correlation Coefficients for Curve Fitting

Curves	PEARL		GOMPERTZ		FLOYD	
Expressions	$Y = \frac{L}{1 + ae^{-bt}}$		$Y = L\, e^{-ae^{-bt}}$		$\frac{F}{F - f} + \ln \left(\frac{f}{F - f}\right) = a + bt$	
Parameters	a =	b =	a =	b =	a =	b =
Coal	1736.7	0.1178	7.502	0.0188	- 6.460	0.1180
Oil	379.6	0.1484	6.109	0.0342	- 4.946	0.1510
Gas	1048.8	0.1550	7.110	0.0296	- 5.959	0.1560
Nuclear	40060.0	0.1426	10.657	0.0157	- 9.598	0.1426
Correlation Coefficients						
Coal	0.9363		0.9600		0.9366	
Oil	0.9633		0.9889		0.9653	
Gas	0.9653		0.9881		0.9661	
Nuclear	0.9587		0.9740		0.9587	

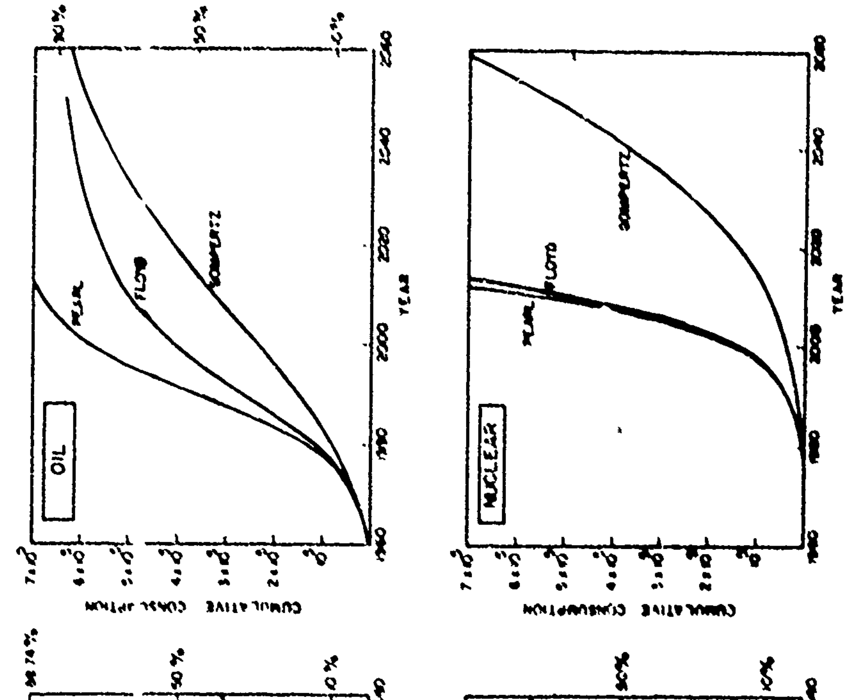


Fig. 2 Pearl, Gompertz and Floyd Curves Fitted to Energy Consumption Data of Table 1

One very simple method of analysis suggested by Sharif and Kabir [17] is to see the behavior of the forecasting model by varying the data extent. Figures 3 to 6 show the projections for cummulative consumption of Coal, Oil, Gas and Nuclear energy using the three growth curves - Pearl, Gompertz and Floyd - but varying the point of forecast from 1964 to 1969 and then to 1974. By looking into these figures as well as Figure 2, the following observations are made:

(i) With the availability of additional data, the steepness of each of the curves in each of the cases decreased consistently.

(ii) Among the different curves, there is a definite trend in relation to overestimation and underestimation for each of the cases.

The above factors can be used as additional information for further evaluation of the available techniques which have been found statistically reliable. However, the answer is not clear cut. Ideally, forecasts should be unbiased and efficient; that is, their average error should be approximately zero (no systematic underestimation or overestimation), and there should be no significant correlation between errors and forecast values. Keeping this in mind, a new model is proposed in the next section.

4. A NEW MODEL FOR ENERGY CONSUMPTION FORECASTING

The proposed new model is based on the following assumptions - (i) the reserves of nonrenewable primary energy resources are finite, and decreasing in time with more consumption; (ii) past consumptions have been increasing exponentially, and there is a positive influence on this trend; and (iii) there is a tendency towards rising resource consciousness, and depletion of reserves create pressure which suppresses the growth in consumption. Thus, it is postulated that the growth in energy consumption is a function of the projected consumption trend and the remaining reserve of the particular energy source.

Empirically, the model can be formulated as:

$$Y_t = [C_t]^{X_{t-1}} \qquad (7)$$

where, Y_t = Estimated consumption for the year t

C_t = Projected consumption for the year t

and X_{t-1} = Resource depletion factor (which is a function of remaining reserve at the end of the year t-1).

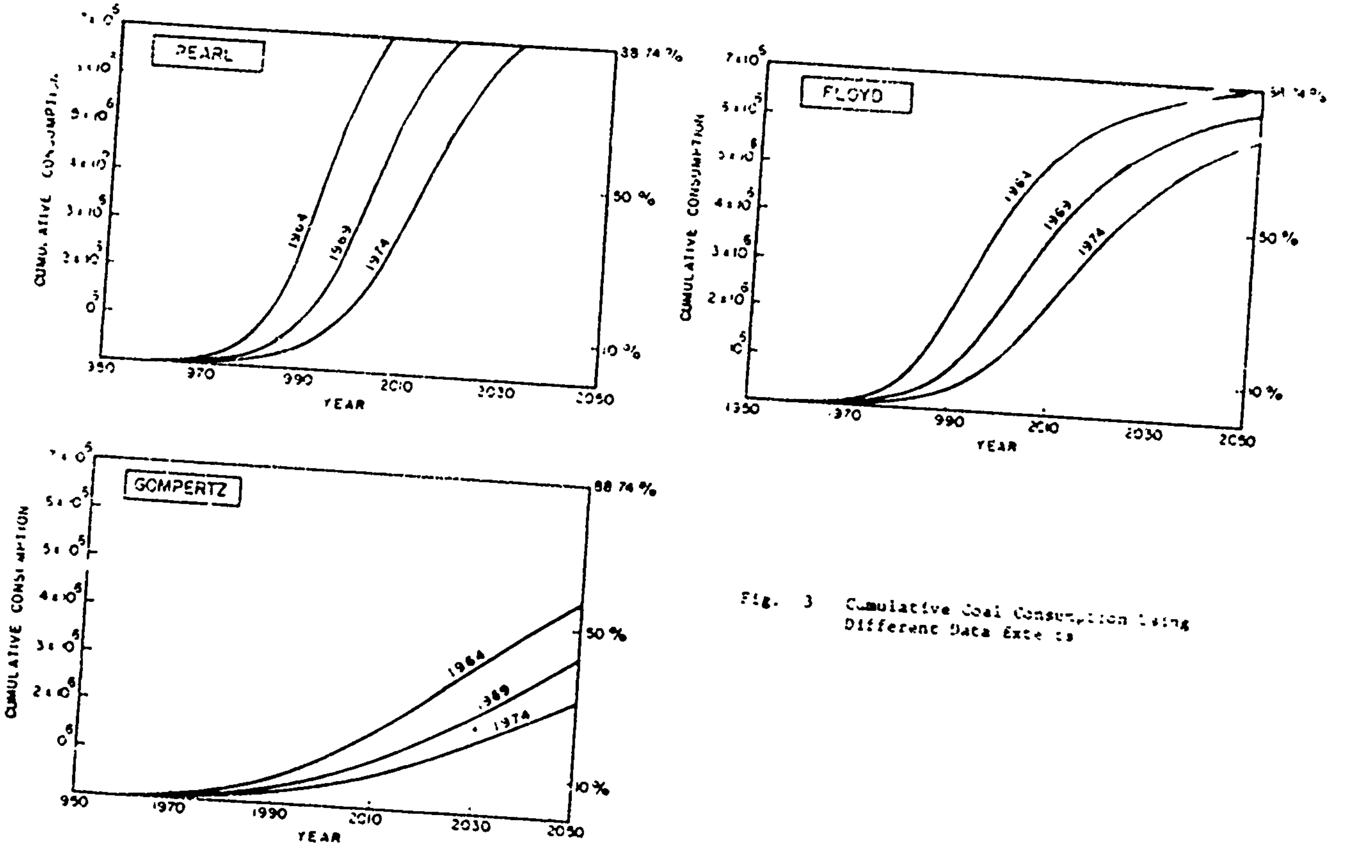


Fig. 3 Cumulative Coal Consumption Using Different Data Extents

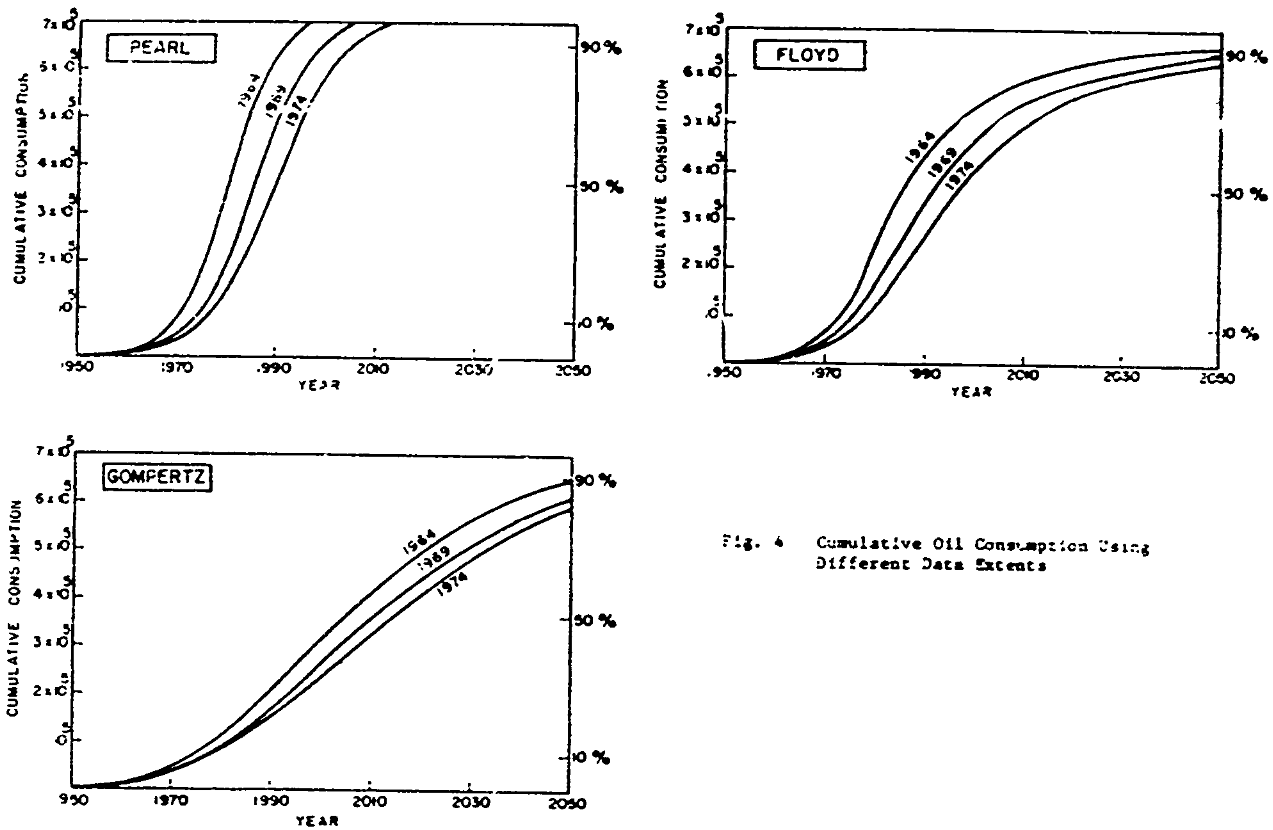
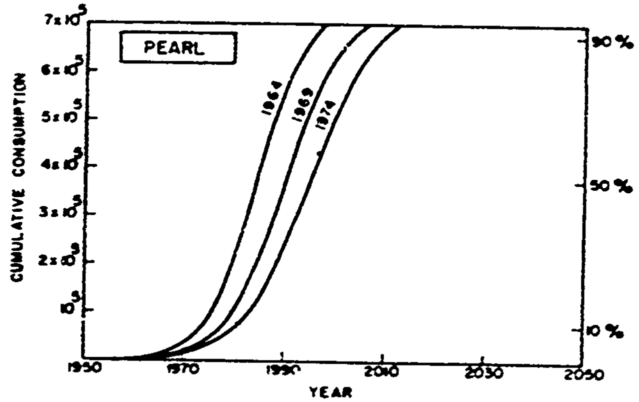
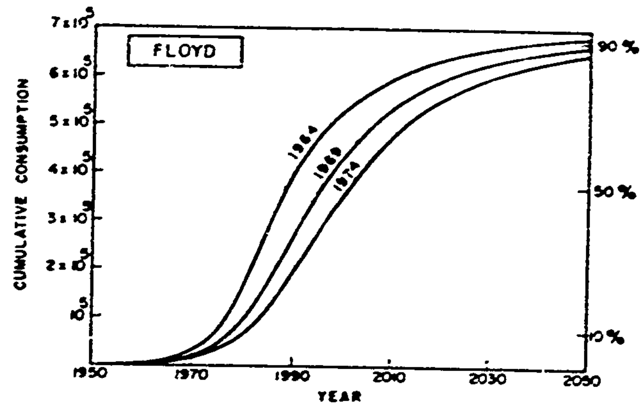


Fig. 4 Cumulative Oil Consumption Using Different Data Extents





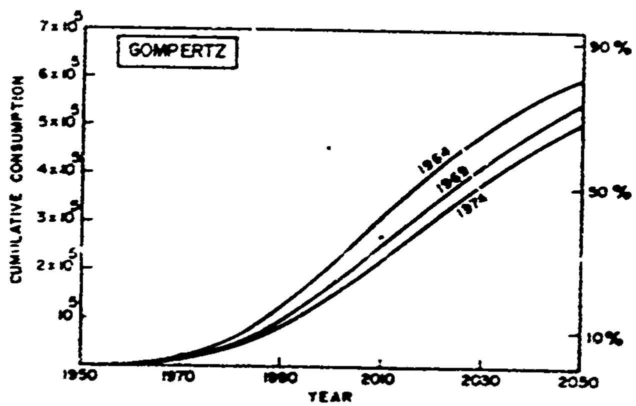


Fig. 5 Cumulative Gas Consumption Using Different Data Extents

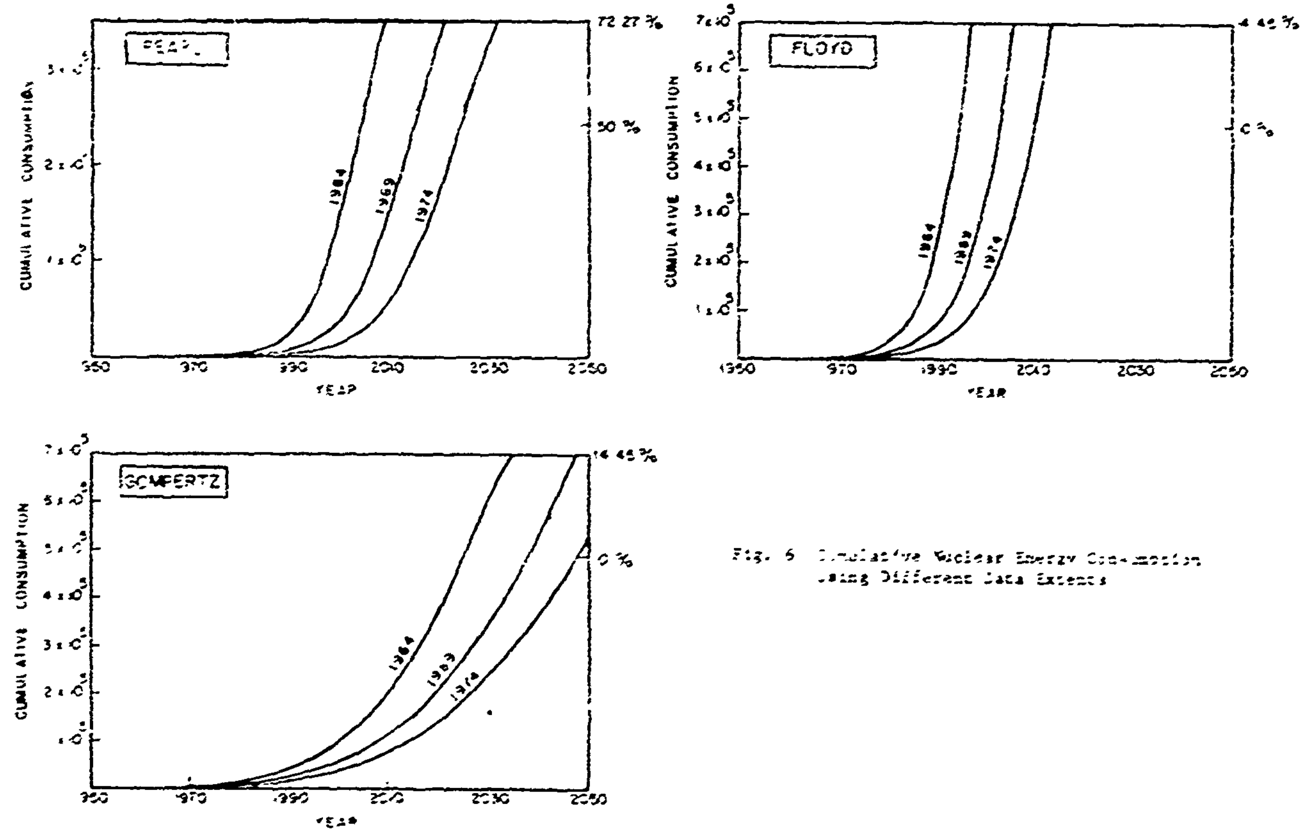


Fig. 6 Cumulative Nuclear Energy Consumption Using Different Data Extents

The projected consumption for the year t can be very easily computed on the basis of our exponential growth assumption:

$$C_t = a\, b^t \tag{8}$$

where, a and b are constants.

The resource depletion factor can be computed as the ratio of the remaining reserve to the initial reserve:

$$X_{t-1} = \frac{R_o - \sum_{j=1}^{t-1} Y_j}{R_o} \tag{9}$$

where, R_o is the initial reserve and

$\sum_{j=1}^{t-1} Y_j$ is the cummulative consumption at the end of year t-1.

Let $r_t = \sum_{j=1}^{t-1} Y_j / R_o$

Therefore, $X_{t-1} = 1 - r_t$ (10)

Now, putting the values of C_t and X_{t-1} in Eq. 7, we get the final expression of the model as

$$Y_t = a^{(1 - r_t)}\, b^{t(1 - r_t)} \tag{11}$$

Using the model represented by Eq. 11, projections are made to describe the consumption patterns of Coal, Oil, Gas and Nuclear energy by considering different data extents (as was done in case of Pearl, Gompertz and Floyd models). The results are shown in Figure 7. The estimates for the parameters and the computed correlation coefficients for the fitted curves using data for the entire period (1950-1974) are shown in Table 3. The correlation coefficients for the proposed model are relatively better than the Pearl, Gompertz and Floyd curves (as can be seen from Table 2). From the projections in Figure 7, one can see that the steepness of the growth curves do not always decrease with the addition of recent data points. Moreover, the magnitude of change in the shape of the growth curves obtained by the proposed model are also relatively smaller compared to those presented in the previous section.

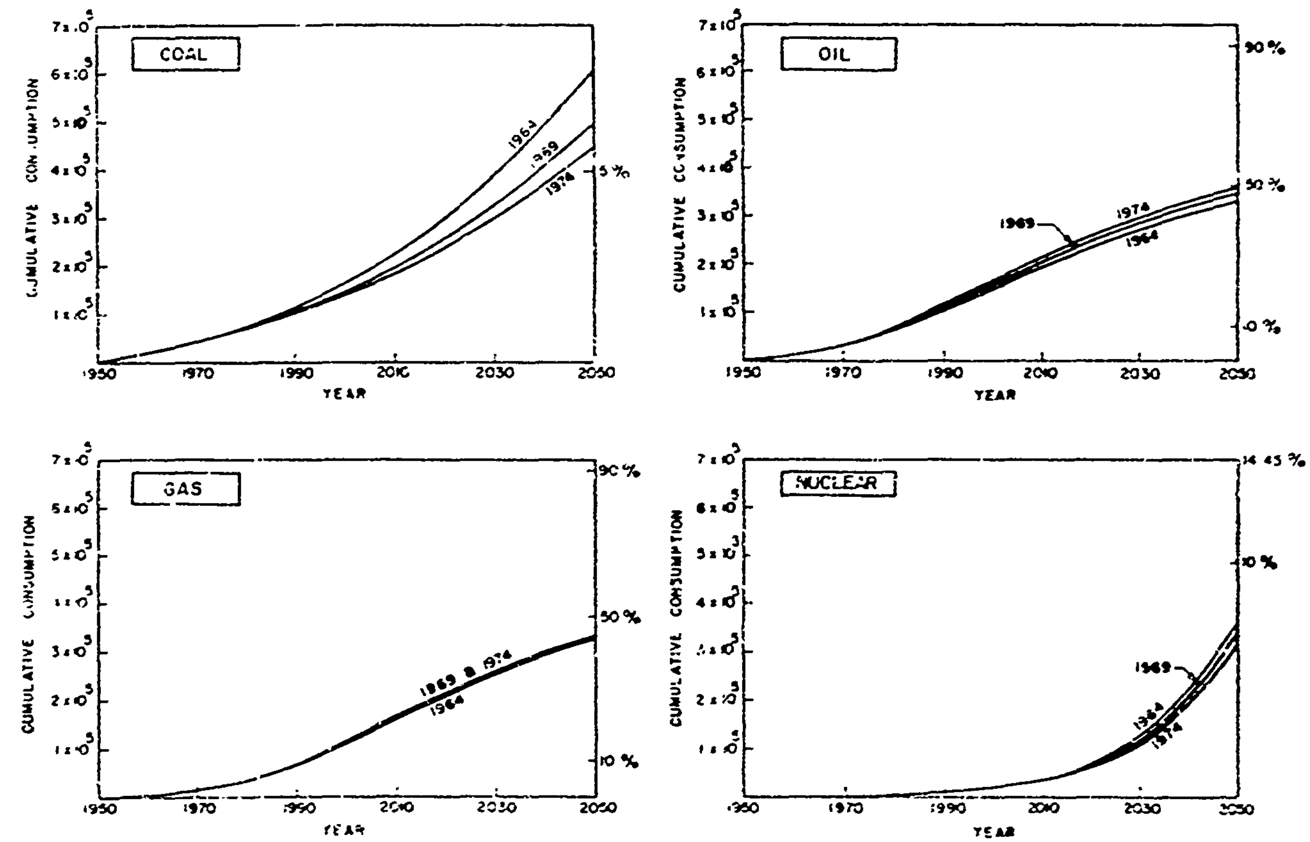


Fig. 7 Proposed Model Projections Using Different Data Extents

Table 3 Correlation Coefficients and Parameter Values for the Proposed Model

Model Expression	$Y_t = a^{1-r_t} b^{t(1-r_t)}$	
Parameters	a =	b =
Coal	1628.87	1.0218
Oil	632.27	1.0957
Gas	250.49	1.0933
Nuclear	42.93	1.0671
Correlation Coefficients		
Coal	0.9715	
Oil	0.9972	
Gas	0.9993	
Nuclear	0.9990	

5. COMPARISONS AND CONCLUSION

This section presents a synthesis of forecasts made by different methods. Complete forecasts regarding the consumption patterns of Coal, Oil, Gas and Nuclear energy using the Pearl, Gompertz, Floyd and the proposed Model are given in Figure 8. For every case all available data (1950-1974) are used. The estimates of the parameters and the correlation coefficients are already given in Tables 2 and 3. Moreover, a comparative forecast regarding the market share situation for the four sources of energy considered in this study is also given in Figure 9. The market share of each energy source has been computed as follows:

$$M_{it} = Y_{it}/T_t \qquad (12)$$

where, M_{it} = Market share of resource i for the year t

Y_{it} = Consumption of resource i for the year t

i = Coal, Oil, Gas and Nuclear energy

and $T_t = \Sigma\, Y_{it}$ for all i .

The forecasts presented in Figures 8 and 9 are now considered vis-a-vis other forecasts and observations available in the literature.

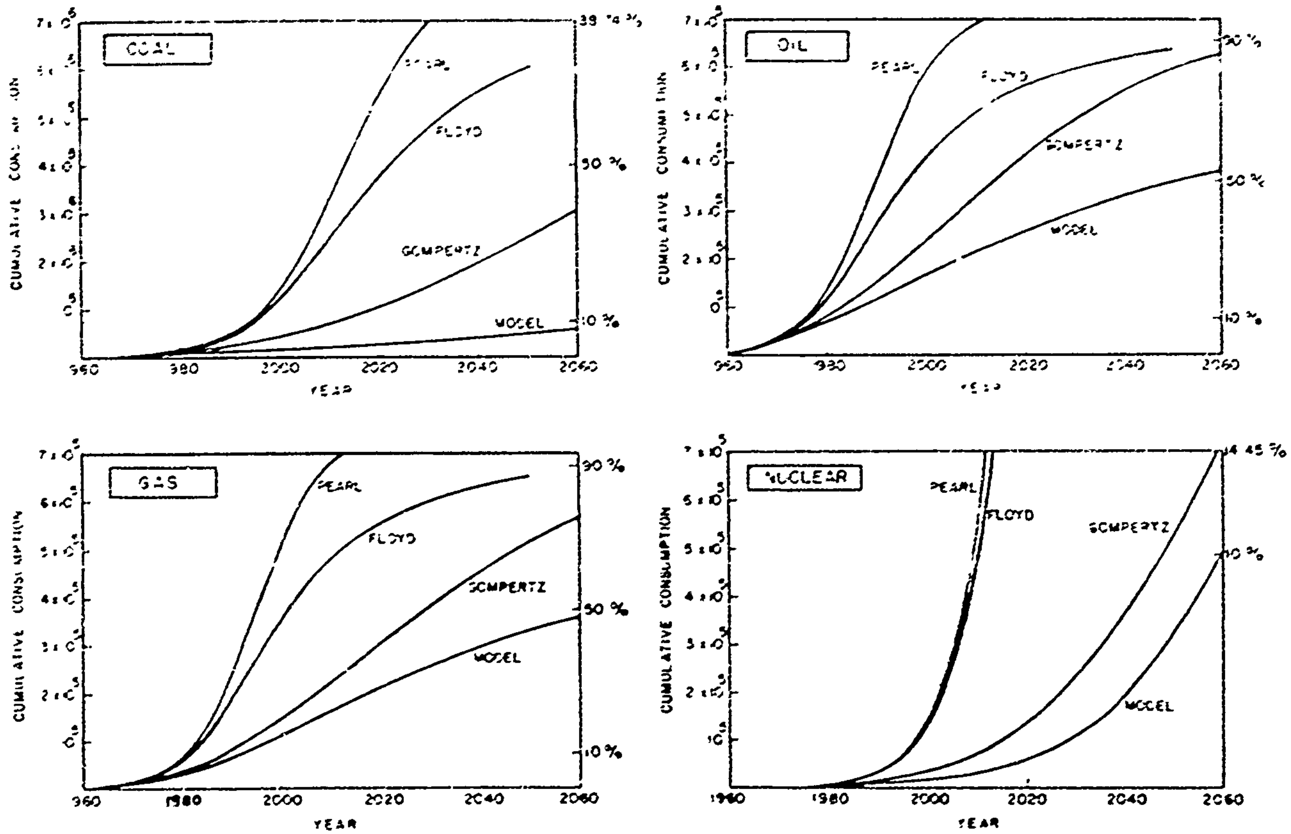


Fig. 3 Comparative Forecasts of Nonrenewable Energy Consumption Patterns

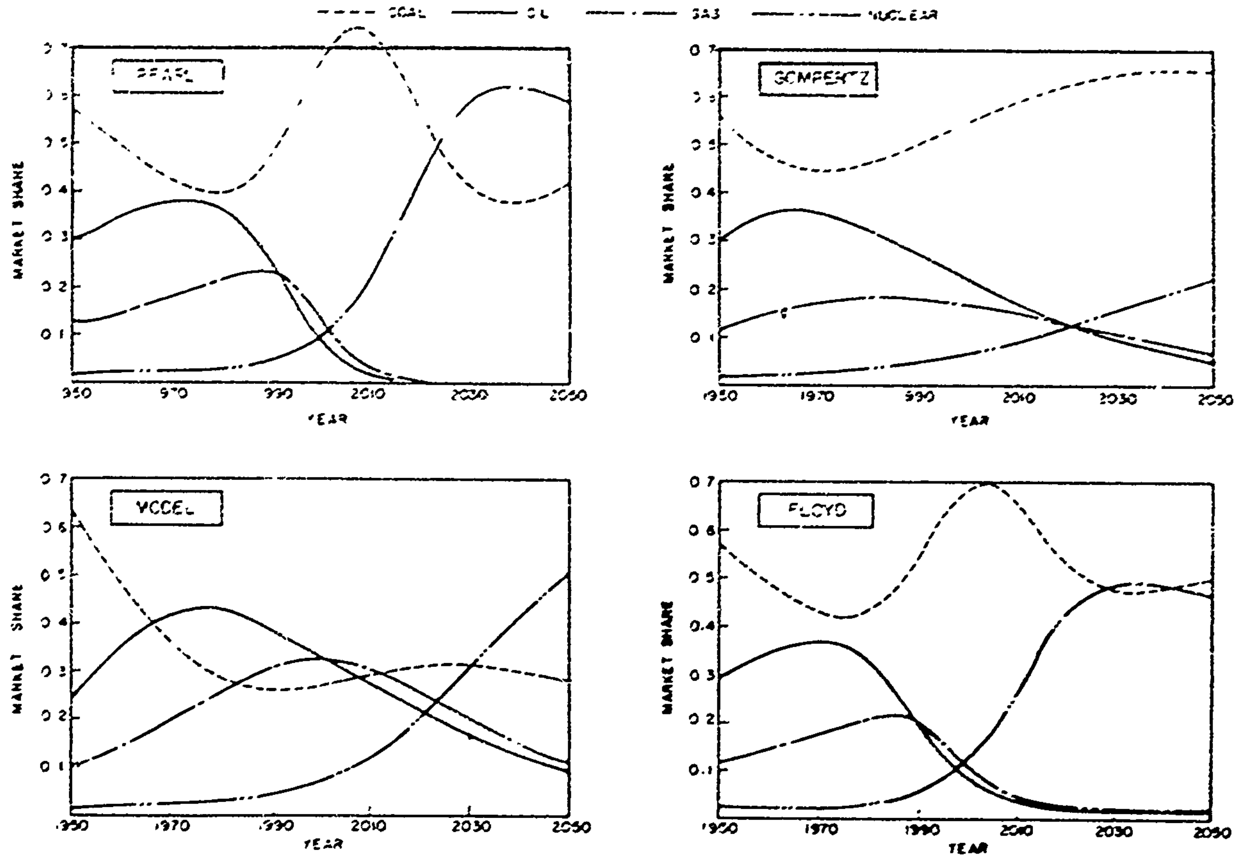


Fig. 9 Comparative Forecasts of Market Shares

Nobody can tell for certain but many specialists cite the figures of M. King Hubbert, a geophysist with the US Geological Survey, who predicts that 90% of all world oil and gas will be gone by 2035, and about 90% of all coal by 2300 [5, 6]. These figures are closer to the forecasts made by the Floyd curve.

If world demand of oil increases at the rate of 2.5% per year (well below the present rate, as a result of price increases) - it will outrun the rate of production sometimes between 1990 and 2010 [14]. This date refers to the inflection point of the growth curves. From Figure 8 it can be seen that, in case of oil consumption, the inflection points for the Floyd curve (1988), Gompertz curve (2005) and the proposed Model (1995) are within this forecasted range.

The Policy Study Group of the MIT Energy Laboratory made the following forecast [15]. The average growth rate in world petrolium demand would increase from 55.7 million barrels per day in 1973 to 92.5 million barrels per day in 1980 if the pace of growth were to continue at 7.5% per year. But this neglects the effect of the recent four-fold increase in oil prices. On the conservative assumption that the price elasticity of demand is - 0.15, the price increase would reduce the 1980 demand by about 19% to 74.9 million barrels per day (which is equal to 4915 million metric tons of coal equivalent). This figure is very close to the forecast made by the proposed Model for oil consumption in 1980 (4443 mmtce.).

Regarding the market share, it has been forecasted that the oil's share of the total energy market can be expected to increase through the end of this decade and then to decline after 1980 [19]. This conforms with the market share forecasts made by the proposed model as shown in Figure 9.

In conclusion, it may be said that although there is no one best model for forecasting energy consumption patterns, this paper shows that some are better than the others.

6. REFERENCES

[1] Austin, A.L. and Brewer, J.W., WORLD POPULATION GROWTH AND RELATED TECHNICAL PROBLEMS, Technological Forecasting and Social Change 3, PP. 23-49, 1971

[2] Bailey, R., TRADITIONAL ENERGY RESOURCES: PRESENT STATE AND FUTURE DEVELOPMENT, Futures 4, NO. 2, PP. 103-114, 1972

[3] Bloom, M.F., TIME DEPENDENT EVENT CROSS IMPACT ANALYSIS: RESULTS FROM A NEW MODEL, Technological Forecasting and Social Change 10, PP. 181-201, 1977

[4] Bossert, R.W., THE LOGISTIC GROWTH CURVE REVIEWED, PROGRAMMED AND APPLIED TO ELECTRIC UTILITY FORECASTING, Technological Forecasting and Social Change 10, PP. 357-368, 1977

[5] Cook, E., ENERGY FOR MILLENIUM THREE, Technology Review 75, NO. 2, PP. 16-23, 1972

[6] ENERGY CRISIS: WE ARE RUNNING OUT, Time, June 12, PP. 33-36, 1972

[7] ENERGY INDEX - 75, Environmental Information Centre Inc., New York, U.S.A., PP. 48

[8] Floyd, A., A METHODOLOGY FOR TREND - FORECASTING OF FIGURES OF MERIT, Technological Forecasting for Industry and Government: Methods and Applications (Bright Ed.), Prentice Hall, Englewood Cliff, New Jersey, PP. 95-109, 1968

[9] Gress, J.V., Hossel, C.H. and Richardson, J.T., MATHEMATICAL TREND CURVES: AN AID TO FORECASTING, Oliver and Boyd Ltd., Edinburg, 1967

[10] Jantsch, E., TECHNOLOGICAL FORECASTING IN PERSPECTIVE OECD, Paris, 1967

[11] Lakhani, H., DIFFUSION OF ENVIRONMENT SAVING TECHNOLOGICAL CHANGE - A PETROLIUM REFINERY CASE STUDY, Technological Forecasting and Social Change 7, PP. 35-55, 1975

[12] Marchetti, C., PRIMARY ENERGY SUBSTITUTION MODELS: ON THE INTERACTION BETWEEN ENERGY AND SOCIETY, Technological Forecasting and Social Change 10, PP. 345-356, 1977

[13] Martino, J.P., TECHNOLOGICAL FORECASTING FOR DECISION MAKING, American Elsevier Publishing Company Inc., New York, 1975

[14] Nutty, P., WHEN WE WILL START RUNNING OUT OF OIL, Fortune, PP. 246-250, October 1977

[15] Policy Study Group of the MIT Energy Laboratory, ENERGY SELF SUFFICIENCY: AN ECONOMIC EVALUATION, Technology Review 76, NO. 6, PP. 23-58, 1974

[16] Sharif, M.N. and Gazi, A.U., A PROCEDURE OF ADAPTING TECHNOLOGICAL FORECASTING MODELS, Technological Forecasting and Social Change 7, PP. 99-106, 1975

[17] Sharif, M.N. and Kabir, C., A GENERALIZED MODEL FOR FORECASTING TECHNOLOGICAL SUBSTITUTION, Technological Forecasting and Social Change 8, PP. 353-364, 1976

[18] STUDY OF THE FUTURES: AN AGENDA FOR RESEARCH, National Science Foundation, Washington, D.C., 1976

[19] SUPPLYING GASOLINE FOR TOMORROW'S CAR, Automotive Engineering 85, NO. 9, PP. 46-51, 1977

[20] Watts, R.G. and Hrubecky, H.F., ON THE LIMITS TO ENERGY GROWTH, Technological Forecasting and Social Change 7, PP. 371-378, 1975

[21] Wessels, T., Blomer, K.H. and Kehse, J., PRESENT STATE OF METHODS OF FORECASTING DEMAND FOR ENERGY, World Power Conference, Tokyo Sectional Meeting, October 1966

[22] White, C.D., ENERGY, THE ECONOMY AND THE ENVIRONMENT, Technology Review 74, NO. 1, PP. 18-31, 1971

[23] WORLD ENERGY SUPPLIES 1950-1974, U.N. Statistical Office, U.N. Publication, New York, 1975

THE FORECASTING MODEL FOR KOREA ELECTRIC POWER CONSUMPTION

SUKHO KANG, Ph. D.

Dept. of Industrial Engineering
Seoul National University
Seoul, Korea.

ABSTRACT. This research will present a case study illustrating the development of a forecasting model for Korea electric power consumption. This research uses a recently developed forecasting technique known as the ARIMA (Autoregressive Integrated Moving Average) technique. The model determination stage, called identification, is followed by parameter estimation, and then diagnostic checking to determine if the model provides an adequate description of the data.

Using the basic structure of the Box-Jenkins model, we identify the specific model that describe the behavior of a time series; from January 1961 to December 1976. After forecasting model development is completed, forecasting for the Korea electric power consumption in 1976 and 1977 is shown and compared with the actual data.

1. INTRODUCTION

At the heart of any planning program is the requirement for suitable forecasts of future conditions. Since forecasts are a major input to decision making, good forecasting is a necessity.

A number of quantitative forecasting techniques are available for making predictions based upon historical analysis: regression, Winters' Three Parameter Exponential Smoothing, and Brown's Discounted Least Squares. Recently Box and Jenkins (1) proposed a structured approach to forecasting, by unifying material and techniques that have been available for a long time into a philosophy for their application.

The methodology suggested by Box and Jenkins represents a systematic approach to modeling and forecasting discrete time series. There are two basic reasons why this methodology can lead to better forecasts than other statistical forecasting methods and thus should be preferred to them. First, using traditional approaches required that the analyst use his own experience as a basis for model development. In the past, experienced analysts have developed methods for identifying the presence of trends and seasonal patterns. Based upon this analysis, a specific model would be defined. On the other hand, using the Box-Jenkins technique, the analyst does not arbitrarily pick a specific model, but instead eliminates inappropriate models until he is left with the most suitable model. Second, the specific form of a given model which is to be used has traditionally been the result of a trial-and-error procedure. Box and Jenkins, however, present a rational structured approach to the determination of a specific model. It should be pointed out that even though this method provides for systematic modeling of time series, management control should be present. All time series techniques are based upon the fundamental assumption that past patterns will repeat in the future. In a dynamic world this is not always the case.

This paper presents a case study using Korea Electric Power Consumption. The historical patterns of power consumption are regular with minimal random variation. This allows us the opportunity to develop the forecasting model for Korea Electric Power Consumption. The Korea electric power consumption data shown in Fig. 1 were selected. Notice that an upward trend is present in Fig. 1.

Furthermore, there is a seasonal pattern in the time series with a cycle length of twelve months. Finally there are random movements away from the underlying trend and seasonal pattern, indicating a stochastic element in the environment.

Table 1 contains the listing of kilowatt consumption for 1961 to 1976 as compiled by the Korea Electric Company. This time series will be divided into two parts. Part one, from 1961 to 1975, will be used to identify, estimate, and diagnostically check a set of time series models during the development phase. The second part, 1976, will be used to evaluate the forecasting model that is developed. Forecasts of 1976 and 1977 will be made with the model and compared against the actual observations.

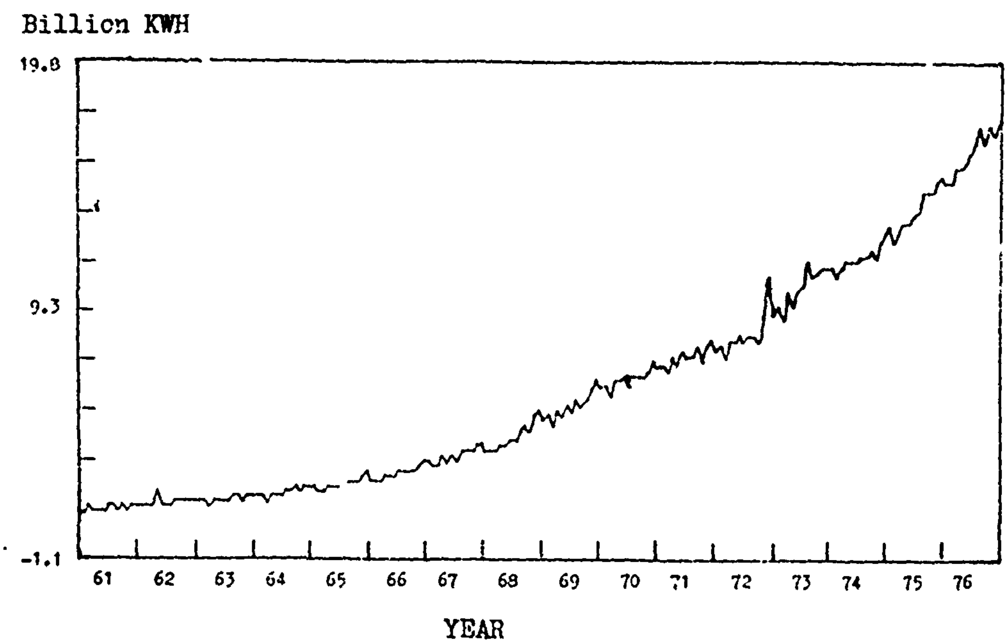


Fig. 1. Plot of Korea Electric Power consumption data.

Table 1. Listing of the Korea power consumption (0000 kilowatts) data.

Year	January	February	March	April	May	June
1961	9291	10812	9471	9347	9487	10600
1962	11650	11639	11281	17023	11043	11866
1963	14023	13028	12261	13335	13666	14494
1964	15694	15199	14111	15592	15614	16569
1965	18876	17594	17058	19399	18984	19819
1966	21901	22861	22425	24413	23920	24877
1967	29317	29202	28231	32026	30915	32659
1968	35224	35047	35337	37342	36678	38183
1969	47793	48725	45603	51666	48236	52442
1970	61575	61133	58453	63529	63081	66094
1971	69193	70752	67098	74281	72030	75605
1972	76995	78244	74526	80866	79866	81855
1973	91554	94935	89455	101522	96391	101628
1974	111954	107692	112277	114247	115314	113982
1975	130186	124678	129100	133215	132673	135900
1976	150898	151030	155972	150488	157408	162962
Year	**July**	**August**	**September**	**Oct.**	**November**	**December**
1961	11230	9434	10375	9823	10548	11508
1962	11641	12495	12933	12617	13903	14212
1963	14183	14797	15059	14332	15011	15372
1964	17711	18085	18931	17496	19429	19910
1965	21195	21555	21312	23142	23485	25777
1966	25251	25642	26288	25172	28132	19966
1967	30255	34175	35240	35144	36832	36291
1968	37928	43021	44683	42860	48305	50305
1969	52202	55090	54262	55558	61215	63215
1970	62490	65503	65926	66395	68305	71305
1971	74412	74730	78111	73238	79037	79868
1972	80138	83336	83826	81010	90206	108360
1973	103772	113813	108676	110351	112426	112160
1974	117824	119219	121057	117932	124408	128896
1975	138240	145012	145451	146573	150367	151638
1976	166766	172679	166006	172979	171563	177275

2. FORECAST MODEL DEVELOPMENT

2.1. Identification-Tentative Models

To identify a tentative model for further investigation, sample autocorrelation functions of various differences of the logged version of the time series for 1961-1976 were calculated and examined. Twelve such sample autocorrelation patterns are shown in Fig. 2, with d at zero, one, two, and three orders of differencing.

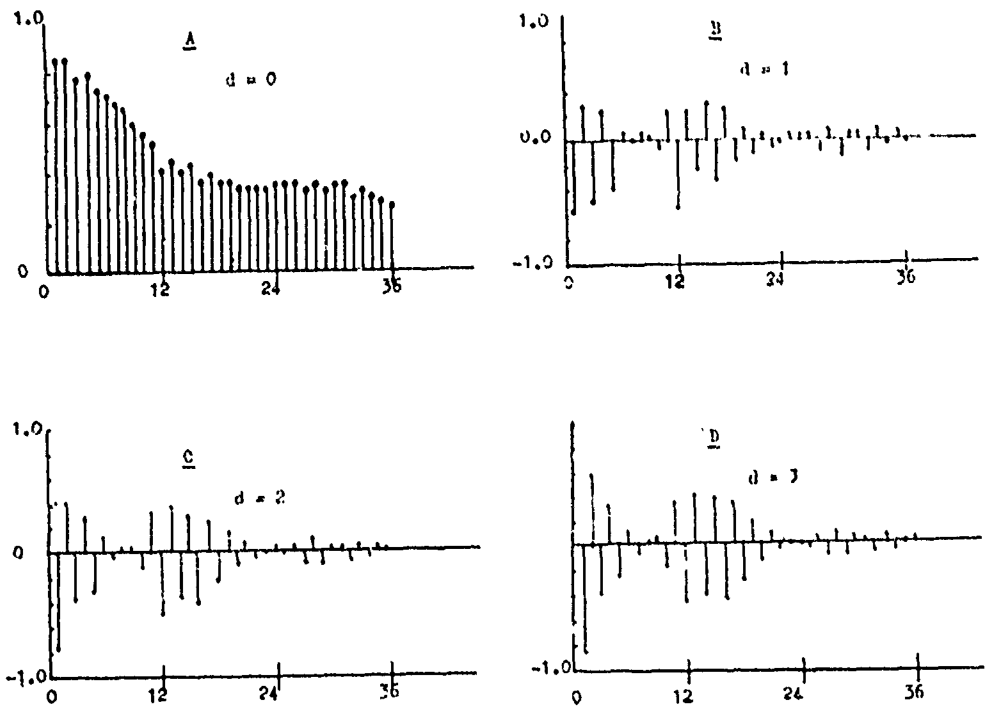


Fig. 2. Sample autocorrelations for differenced series indicated by *d*.

The sample autocorrelations of the logged version of the time series with no differencing are shown in Plot A. The autocorrelations are high for early lags and remain high for increasing lags, i.e., there is no quick damping tendency evident in Plot A. This is also the case shown in Plots B, C, and D for regular first through third differencing of the series. Observe that in checking the partial autocorrelation pattern (Fig. 3), there are large values for the first two lags. Therefore, a tentative model for

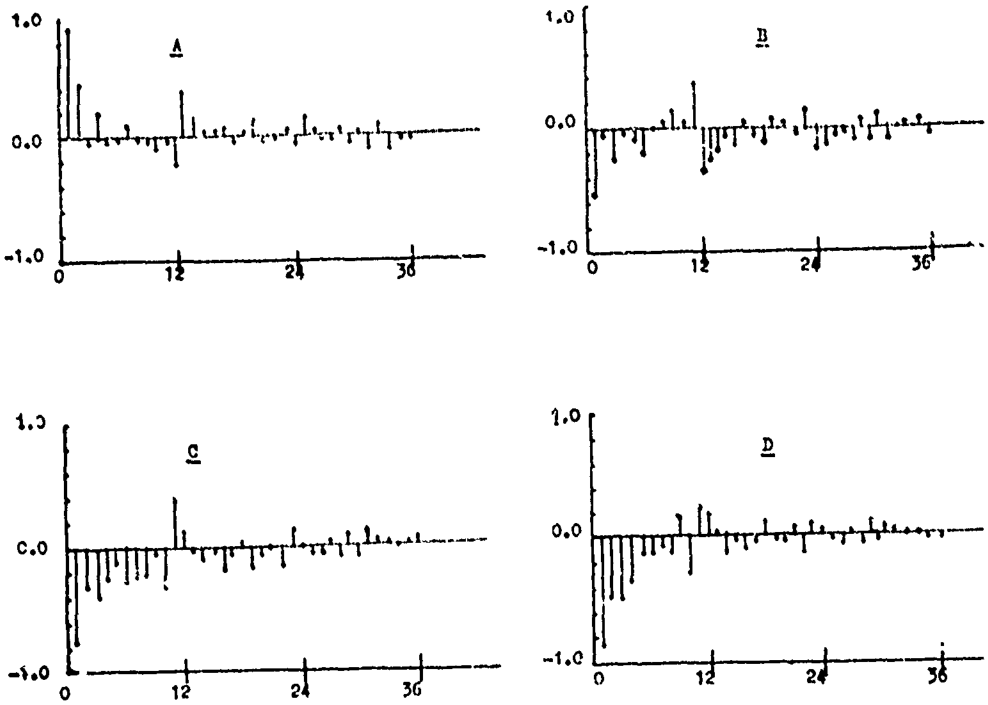


Fig. 3. Partial autocorrelation plots for Korea power consumption differential series.

further investigation is

$$(1 - \phi_1 B)(1 - \phi_1' B - \phi_2' B^2)(1 - B^{12}) Z_t = \theta_0 \quad (1)$$

where Z_t is the logged series, $(1 - B^{12})$ is the seasonal difference applied to Z_t to achieve stationarity, $(1 - \phi_1' B - \phi_2' B^2)$ is the second order autoregressive operator applied to the stationary series to reduce it to random shocks, and θ_0 is white noise.

2.2. Estimation - Model (1)

We estimated the model (1) coefficients with the 1961-1976 time series data, and obtained the following coefficient values:

$$\phi_1 = 0.0187, \phi_1' = 0.4379, \phi_2' = 0.5033, \theta_0 = 0.08$$

2.3. Diagnostic Checking - Model (1)

The Chi-square test of the residual sample autocorrelation and examination of Fig. 4 reveals that model (1) is inadequate. A significant autocorrelation at lag 12 indicates that the residuals are not independent, i.e., the $N(0,\sigma_a^2)$ assumption is not satisfied.

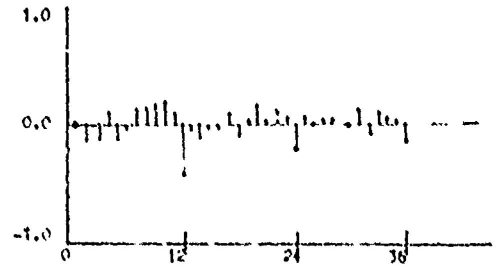


Fig. 4. Sample autocorrelation functions of residuals.

2.4. The Identification of a New Model from the Model (1) Results

The sample autocorrelation pattern for the residuals suggests a way of improving model (1), that is, a new model, (2), is indicated. The large spike at lag 12 indicates the need for introducing a moving average operator involving period 12. The newly identified model can be represented as

$$(1 - \phi_1 B)(1 - \phi'_1 - \phi'_2 B^2)(1 - B^{12})Z_t = (1 - \theta_{12} B^{12})a_t. \quad (2)$$

2.5. Estimation - Model (2)

The estimation process for model (2) produces the following model coefficient values:

$\phi_1 = -0.287$, $\phi'_1 = 0.717$, $\phi'_2 = 0.284$, $\theta_0 = 0.0018$, $\theta_{12} = 0.768$

2.6. Diagnostic Checking - Model (2)

The application of the Chi-square test to the autocorrelations of the residuals at the 95% confidence could not reject the null hypothesis that the residuals are randomly distributed. Therefore, model (2) can be accepted as a reasonable model of the time series.

3. FORECASTING KOREA ELECTRIC POWER DEMAND

We develop the forecast equation for model (2). This forecasting equation will be used to make monthly forecasts from a single in time (T). Model (2) is rewritten as:

$$(1 + 0.287B)(1 - 0.717B - 0.284B^2)(1 - B^{12})Z_t = 0.0018 + (1 - 0.768B^{12})a_t. \qquad (3)$$

This equation can be expanded and rearranged as

$$Z_t = -0.43Z_{t-1} + 0.49Z_{t-2} + 0.082Z_{t-3} + Z_{t-12} + 0.43Z_{t-13} -0.49Z_{t-14} - 0.082Z_{t-15} + a_t - 0.768a_{t-12} + 0.0018. \qquad (4)$$

We can rewritten (4) as our forecasting function to forecast l periods in the future beginning with time period T + l:

$$Z_t(l) = -0.43Z_{t+l-1} + 0.49Z_{t+l-2} + 0.082Z_{t+l-3} + Z_{t+l-12} + 0.43Z_{t+l-13} - 0.49Z_{t+l-14} - 0.082Z_{t+l-15} + a_{t+l} - 0.768a_{t+l-12} + 0.0018. \qquad (5)$$

With December 1975 as our forecasting origin (T), the forecasts and the associated two standard error limits are listed in Table 2. These data are also plotted in Fig. 5 along with the actual observations for 1976-1977. A comparison of the forecasts with the actual observations shows an average difference of approximately 1.7% between the forecast and actual values, indicating the effectiveness of the model. (See Fig. 5 and Table 3.)

Table 2. Korea power consumption

Period	Forecast of Korea power (0000 kilowatts)			
	LCL	Forecast	UCL	Actual
1976				
January	146179	151377	156575	150900
February	143967	149625	155283	151000
March	143892	150550	157209	155999
April	148750	156064	163379	156500
May	147444	155430	163416	157400
June	149674	158258	166842	162999
July	157610	162808	168006	166800
August	162592	168250	173909	172700
September	162062	168720	175378	166000
October	161235	168550	175576	173000
November	165604	173591	181576	171600
December	169061	177645	186230	177300
1977				
January	169846	175044	180242	
February	169018	174676	180334	
March	169723	176382	183040	
April	173659	180974	188288	
May	172810	180796	188781	
June	175830	184414	192998	
July	177436	186587	195739	
August	182011	191695	201380	
September	180666	190858	201049	
October	181729	192404	203079	
November	184974	196113	207252	
December	189090	200674	212259	

LCL = Lower Confidence Limit
UCL = Upper Confidence Limit

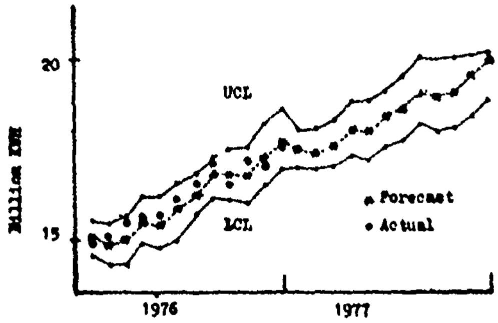


Fig. 5 Forecast Korea power consumption

Table 3. Korea power consumption

	Absolute Percentage error of forecast
	1976
January	0.3%
February	0.9
March	3.5
April	0.3
May	1.3
June	2.9
July	2.4
August	2.6
September	1.6
October	2.6
November	1.2
December	0.2
Average error per month	1.7%

4. CONCLUSION

This paper presented a case study illustrating the development of a forecasting model for Korea power consumption. The basis for model development is the three step Box-Jenkins procedure of identification, estimation, and diagnostic checking. Once development is completed, forecasting from single or multiple origins can be done. For effective and efficient forecasting, management must take an active part. It means that historical data only provide a starting point. Thus management experience and judgement must be involved.

REFERENCES

(1) Box, G. E. P., and Jenkins, G. M., TIME SERIES ANALYSIS, Holden Day, 1970.

(2) Bedworth, D. D., INDUSTRIAL SYSTEMS, The Donald Press Co., 1973.

(3) Brown, R. G., SMOOTHING FORECASTING AND PREDICTION OF DISCRETE TIME SERIES, Prentice-Hall, 1963.

(4) Kang, S., THE USE OF ARIMA TECHNIQUE TO THE DEVELOPMENT OF COMPUTER FORECASTING MODEL, Engr. Report, Seoul National University, Vol. 10, NO 2, Oct. 1978.

(5) Mabert, V. A., AN INTRODUCTION TO SHORT TERM FORECASTING USING THE BOX-JENKINS METHODOLOGY, AIIE, 1975.

(6) Pack, D. J., Goodman M. L., and Miller R. B., COMPUTER PROGRAMS FOR THE ANALYSIS OF UNIVARIATE TIME SERIES USING THE METHODS OF BOX AND JENKINS, Technical Report, NO 296, University of Wisconsin, April 1972.

ON ENERGY SYSTEM MODELING
FOR POLICY ANALYSIS

BYONG-HUN AHN

Korea Advanced Institute of Science
Seoul, Korea

ABSTRACT: This work provides an introduction to the scope, applications, methodology, and content of energy related engineering-economic models, particularly those developed and used in the United States. Some important methodologies used to implement these models are briefly discussed, a classification of models is provided, and representative models are reviewed. This review is not intended to be exhaustive or to merely list models. Rather, the models are reviewed to illustrate and analyze the structure and trends of recent and current efforts by energy system modeling community to provide useful and constructive analytical tools for understanding and solving energy planning and policy problems.

Since the Arab oil embargo of 1973, many governments have been concerned with the importance of energy in policy-making as a vital component in the economic and social well-being of a nation. The energy problem emerges as a challenge to our society. It calls for the development of innovative energy supply technologies and the improved conservation actions. To meet these needs, there have been initiated numerous research and development activities around the world to improve energy supply and consumption technologies. In equal measure, however, the need is for improved decision making to choose among the many available energy futures: a large part of the energy decisions is to define what the critical issue is, and then to examine and evaluate the difficult tradeoffs between the economic, societal and political costs and benefits associated with alternative energy futures.

The national energy system is highly interdependent. A variety of energy forms can be used to satisfy many ultimate requirements, alternative production technologies compete for many of the same resources, physical and technological limitations can restrict the supply or potential uses of energy, environmental effects or other externalities can alter the permissable production and use patterns, and new technologies can rapidly transform the available options.

The development of a national energy policy must include a description and evaluation of these and other complex interdependences. The evaluation of alternatives must be done within the framework of a flexible system able to describe and quantify the many components of the energy system.

Often the divergence in the proposals for new action results from simple but fundamental differences in views about the nature of the problems and the proper solutions. If made explicit, these different views can be compared and evaluated on the rather common conceptual basis. This is where formal policy models can contribute by providing a common framework and capability for organizing and extending the necessary investigation of energy alternatives. This motivation for developing energy policy models has been well bought since 1973 event in particular.

Energy policy models are formulated using theoretical and analytical methods from several disciplines including engineering, economics, operation research, and management science. Techniques of applied mathematics and statistics used to implement these models include mathematical programming, econometrics and related methods of statistical analysis, and decision analysis.

The purpose of this paper is to provide an introduction to the scope, application, methodology, and conceptual contents of selective energy policy models, particularly those introduced after the 1973 oil embargo. The important methodologies used to implement these models are surveyed, a classification of models based on the level of aggregation is discussed, and representative models in each category are reviewed. The coverage is not intended to be exhaustive. Rather, this paper illustrates the structure of recent and current efforts by energy policy modelers to provide useful and comprehensive tools for understanding and discussing important energy planning and policy problems.

2. METHODOLOGIES

Energy system models are formulated and implemented by the theoretical and analytical methods of various disciplines including engineering, economics, and operations research/management science. A particular model frequently employs more than one method, and it is often difficult to classify a model in a clearcut manner. Models based primarily on economic theory tend to emphasize behavioral characteristics of decisions of producers and consumers, while models based on engineering concepts tend to be concerned with the technical and physical aspects of production and consumption processes. A typical example of models combining these two approaches is the Project Independence Evaluation System (PIES) model of the US Department of Energy [1] . This model is an evidence of growing trends in energy modeling in order to provide a more comprehensive framework in which to forecast the condition of future energy markets under alternative assumptions concerning emergence of new production, conversion, and consumption technologies. This trends illustrates the importance of an explicit recognition of technical constraints in formulating and evaluating alternative national or industrial energy policies or strategies.

Models based on aforementioned concepts are formulated and implemented by such typical tools as mathematical programming (in particular linear programming), activity analysis, econometrics, input-output approach, system dynamics and related methods of statistical analysis.

Process models [1,2,3] are usually implemented using the programming techniques associated with network and activity analysis, while the behavioral models employ statistical methods.

Mathematical programming, in particular linear programming, has been used in energy system modeling to capture the technical or engineering details of specific energy production, conversion and utilization processes in a framework that is rich in economic interpretation. Series of activity variables are defined representing the levels of activity in various processes, and

these are constrained through demand requirements, supply constraints, and any other special relationships that must be defined to incorporate technical reality or other physical constraints that must be met. The dual problem formulated in terms of prices is associated with any LP problem formulated in quantities. This price concept associated with a dual problem provides an important link between process and economic analyses. Another characteristic of the optimization approach is that an objective function should be defined to be either minimized or maximized. Sometimes this property creates a conceptual difficulty in choosing a proper single objective criterion. Once an acceptable objective criterion is chosen, however, this approach becomes a predictive model rather than a descriptive one.

Econometrics is another frequently used technique in energy system modeling. It is concerned with the empirical representation and validation of economic theories and laws via the application of statistical techniques to estimate the structural parameters of one or more equations assumed and derived from economic theory. Econometric methods are used in modeling two types of energy processes: behavioral and technical processes. Behavioral processes are typified by various consumer reactions to market signals such as prices, and many studies on market demand forecasts for energy products have employed econometric approaches.

Technical processes can be described via econometric estimation using a proper structural assumption. An example would be the production function of a firm in which maximum potential output is a function of the quantities of inputs available such as capital, labor, energy, and other inputs. Translog cost function (see Jorgenson and Lau [4]) is one of popular and well-received functional forms. Technical relations obtained above could also be used in deriving behavioral relations concerning the firm's demand for input factors, for example energy products, at a given output level.

Interindustry techniques or the input-output approaches are frequently employed in energy policy modeling. Unlike the general input-output tables where coefficients are based on a common unit, that is, the monetary unit, the energy sector input-output coefficients are represented via physical units, such as BTU, kilocalorie or joules. Then, the energy I-O table shows the flow pattern of energy products through the components of the economy. Important and often critical assumptions of this approach include fixed technology and zero price elasticity (due to the assumed independence of input proportions with respect to relative prices). With these assumptions, the I-O approach can be limited to the investigation of direct and indirect energy requirements and effects (interdependences among energy technologies) for a given level of final energy product demands. To overcome

this limitation, Hudson and Jorgenson [5] has developed an energy policy model where the energy I-O eoefficients are represented as a function of relative prices of all inputs. This important development will significantly increase the potential of I-O approach in energy system modeling.

In many applications, the methodologies mentioned above are employed jointly rather than individually. For example, the mathematical programming representation of the energy supply sector utilizes the energy I-O concepts in describing the energy technologies. Some models, which are concerned with market equilibria, utilize the mathematical p ogramming approach to describe the supply side and the econometric methods to represent the consumption behavior within a single modeling framework.

3. Classification of Energy Models

The energy policy models are discussed in several groups according to their scope, and they range from models of a single fuel to models covering full energy system linked to the rest of the economy:

1) single fuel or related fuels market models, which concerns with supply, demand and balancing relationships for individual or related fuels,
2) energy system models, which models supply and demand relationships for all energy related activities, but does not explicitly consider the feedback phenomena of the energy sector upon the rest of the economy, and
3) energy-economic models, which includes the two way linkage between the energy sector and the rest of the economy.

In discussing and comparing models, interested features are the basic concepts used, the methodology implemented and other relevant characteristics. For the supply sector and the demand side, the level of aggregation, the level of interfuel substitution and factor (energy and nonenergy) substitution, dynamics and trends (of exogenously given informations on time profiles of population, labor force, and labor productivity, for example) are among features to be compared between different models. Supply-demand balancing meahanisms differ among the models to be reviewed here, and presents a major source of model diversity. In this article, balancing mechanisms are classified into two major categories, positive and normative. Positive models postulate the existence of certain behavior on the part of consumers(e.g., maximizing utility) and the producers (e.g., maximizing profits) and assume that these sectors communicate through competitive

markets. The primary focus of information exchange in the markets is the distribution of relative prices of the products and resources. For the positive models, balance is achieved when these quantities and prices are equal. The system is then said to be in a market equilibrium. Normative models here start from the same technological description of production and consumption possibilities, but assume that the producing and consuming sectors operate cooperatively to maximize some joint criterion function. The system is in balance when the physical flows match and the production-consumption activities are at levels which maximize the criterion function over the feasible values. The comparison features mentioned above are similar to those used in the model comparison works of Energy Modeling Forum [6] .

4. Industrial Market Models for a Single Energy Form

Models for energy industry markets include process and econometric models as well as integrated process/econometric models, which characterize both supply and demand for a specific or related set of energy products. Typically, models in this category focus on the supply side of the market. Process models are used most often for characterizing energy product supply and capacity expansion. Process-oriented supply models have been developed and applied most extensively to the analysis of oil refining operations and to the capacity expansion plan of the electric utilities. Many models in this category assume the final demands given exogenously, and try to operate in such a manner to minimize the total cost of production while meeting the specified demands. The LP technique is one of the major tools employed for this type of models.

However, in this type of models, consumer responses to various changes in supply side situations (resulting into changes in marginal costs of production, which affect market prices which in turn affects the levels of market demand quantities) cannot be reflected in the final solutions. This could be a serious drawback, when there exists large price elasticities of demands. In this section, we examine models which handles the demands and prices endogenously allowing for the price induced conservation on the part of consumers.

4.1. Kennedy World Oil Market Model [7]

This model is international in scope. It deals with oil production, consumption, trade patterns and pricing. The international price of oil is calculated on the assumption that OPEC will act so as to maximize its net economic return. This differs from many energy models where the international

price of oil is an exogenously specified input parameter.

Kennedy's is a multi-commodity, multi-region single-period economic equilibrium model. This model includes crude oil supply functions and also demand functions for four refined products in each of six regions of the non-Communist world. Cross-price elasticities are taken to be zero. Linearity is assumed for each of the supply and demand functions. Market equilibrium may therefore be computed by solving a quadratic programming problem for maximum net economic benefits. Much like the linear programming models employed routinely for refinery coordination within individual oil companies, this one includes interregional transport and oil refining activities. The costs of these activities are subtracted from gross economic benefits (the areas under individual demand curves) to calculate the "net economic benefit" quadratic objective function.

There are two major methodological limitations in Kennedy's model. First, it is static, and therefore excludes intertemporal phenomena such as resource depletion, and the dynamics of supply and demand responses to higher prices. Second, it does not allow explicitly for competition between oil and alternative fuels. This is a limitation of not only Kennedy's but all single fuel market models, since energy demands can be met by many alternative energy products.

4.2. Baughman-Joskow Electricity Model [8]

There exist numerous models dealing with electricity generation operation and facility expansion problems. The space limitation does not permit to review these models. Rather, in this paper, the Baughman-Joskow model is selected for discussion, for this model covers most of major policy questions associated with electricity utilities. This model combines an engineering supply model with an econometric demand model and links the two with an explicit model of the regulatory process by which the price of electricity is determined.

The supply model for electricity is regional, encompassing the nine regions. Each region is assumed to have eight potential plant types available, and a ninth type, hydro-electric, is treated as exogenous. The plant types are gas turbines and internal combustion, gas-fired thermal, coal-fired thermal, oil-fired thermal, light-water uranium reactors, plutonium recycle reactors, high-temperature gas reactors and liquid-metal fast breeder reactors. The model characterizes the decision process by which operation and expansion of the electricity supply system takes place based on cost minimization techniques used by the industry. The econometric demand model forecasts regionalized electricity demands by the industrial and residential/commercial sectors as functions of the

prices of electricity and alternative fuels -- coal, oil and gas.

This model is not an optimization model designed to calculate an intertemporal economic equilibrium. Here capacity expansion decisions are based upon electricity demand expectations formed by exponentially weighted moving averages of previous demands. In general, these projections will differ from actual future electricity demands, for this model simulates the electric utility's behavior based on a myopic manner -- in contrast to multi-period mathematical programming capacity expansion models, which imply perfect foresight.

Unlike many models in this category, this model has a financial-regulatory submodel simulating the process by which the electric utility industry raises investment capital and sets the price of electricity in accordance with the administrative procedures of regulatory agencies. This submodel allows for the prices faced by consumers to differ from the marginal costs used for operating and capacity expansion decisions within the utility industry. This model was used to assess the future of the US nuclear industry [8], raising serious questions as to the future financial viability of the nuclear equipment manufacturers.

Like the Kennedy's oil market model, this model as a single energy form model cannot accomodate competition between electricity and alternative fuels. These single fuel market models can serve many policy questions, in particular, those of near-term periods. However, due to the high interdependence among fuels, it is often necessary to model inter-fuel competition (substitution) in an explicit manner. This necessitates the simultaneous treatment of components of the full energy system.

5. Energy System Models

The national energy system is highly interdependent. A variety of energy forms can be used to satisfy many ultimate requirements, alternative production technologies compete for many of the same resources, physical and technological limitations can restrict the supply or potential uses of energy. Even though a particular policy question concerns with a specific energy form, it should be analyzed in conjunction with other energy products, recognizing high and complex interdependences among them. Energy system models, in this regard, are generalizations of single fuel market models represented in the previous section. But these models exclude explicit representation of interactions between the energy section and the remainder of the economy.

The energy system models included in the paper are Brookhaven models, PIES model, SRI/Gulf/DFI model and ETA

model.

5.1. Brookhaven Models [3,9,10,11]

Brookhaven National Laboratory of US has developed and is currently extending a series of energy models used to evaluate energy policy and energy research and development alternatives. These models are identified as (1) Brookhaven Energy System Optimization Model (BESOM), (2) Dynamic Energy System Optimization Model (DESOM), and (3) a combination of BESOM with an input-output model developed at the University of Illinois, and (4) a combination of BESOM and a model developed by the Data Resources Incorporated.

BESOM, in essense, is a generalized linear programming transportation model of the Koopmans-Hitchcock-Kantorovich form. It minimizes the total cost of satisfying a given set of energy demands. However, the optimization is performed on the basis of annual cost in a single target year rather than a minimum present worth over some planning horizon. Typically, results are reported for the years 1985 and 2000. This static simplification is probably adequate for the analysis of interfuel substitution. An intertemporal dynamic model is not needed for this range of policy questions.

BESOM is formulated as a 'weighted distribution model'. Given a set of end-use energy resources and conversion efficiencies, the energy flows throught the system are calculated so as to minimize total costs subject to the demand and resource constraints in the target year. Figure 1 illustrates the network structure (Reference Energy System) based on a detailed projection of end-use demands for 1985. Each link in this network represents a physical process or a mix of physical processes for a given activity. Each successive step in the supply chain is integrated along with the end-use devices. The 'sources' in this network are identified with primary energy supplies (e.g., underground coal, imported oil, hydroelectric, nuclear, etc.). The 'destinations' are identified with end-use demands (e.g., air conditioning, space heating, petrochemicals, power, etc.). The costs of extraction, refining and conversion, transportation and storage, and final utilization are assigned to each supply-demand combination. BESOM has been used to study the competition between electric and nonelectric energy forms in specific end-uses, the feasible range of electrification of the energy system, the break-even costs and the optimal implementation model of new energy technologies.

A dynamic, time-phased version of this model has been developed, and is named DESOM. It incorporates the same technical details and constraints as BESOM, but treats plant expansion, resource depletion, and capital requirements

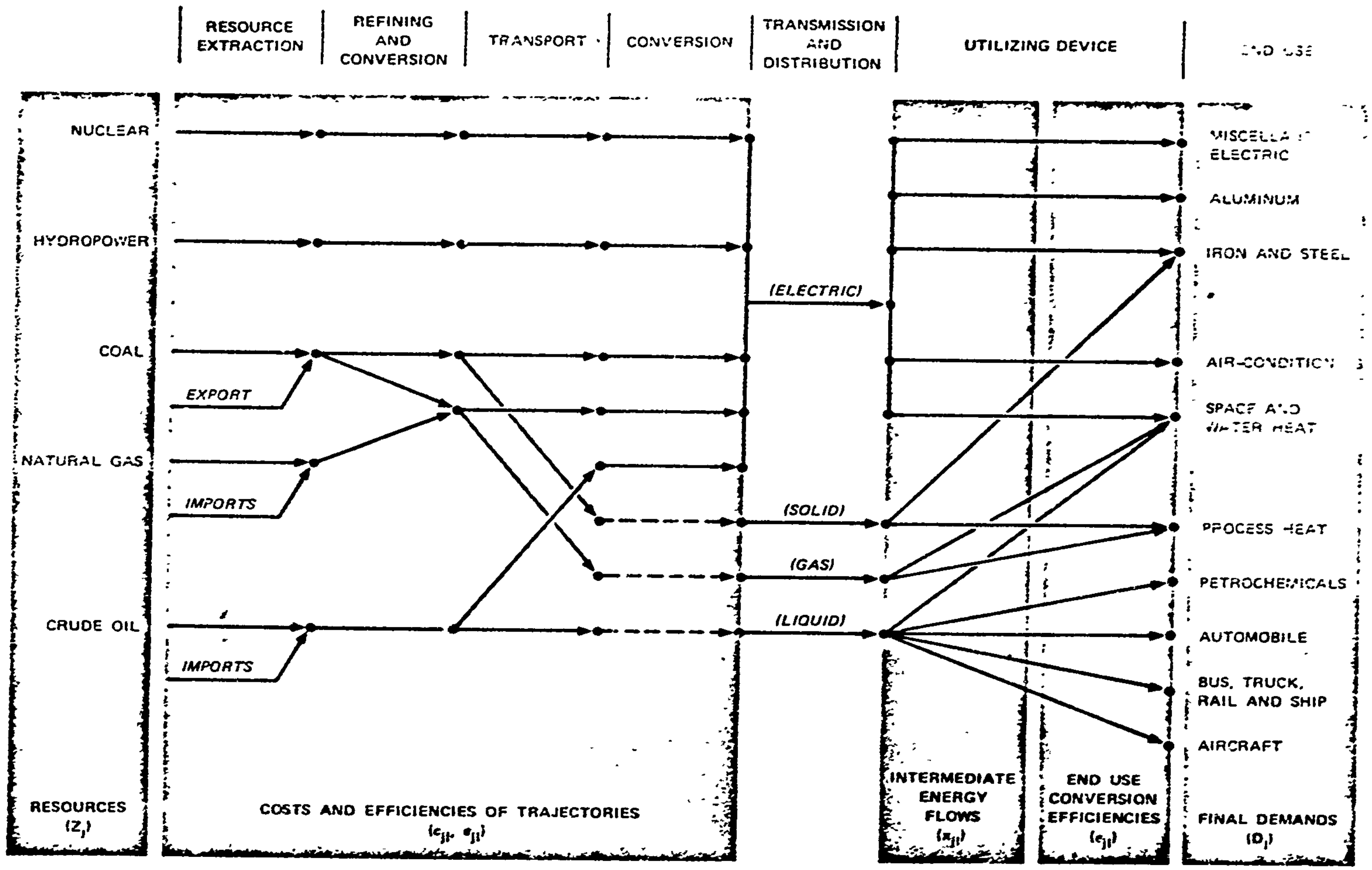


FIGURE I BESOM ENERGY NETWORK

explicitly. Like its static counterpart, DESOM assumes future demands for energy services as given and then determines the combination of energy resources and technologies to be used over time in order to meet those demands at least cost. The assumption of the fixed demand growth could be a critical shortcoming for a certain class of problems, especially where the prices are involved. If the supply system is constrained so that marginal costs rise significantly, these high prices will cause consumers to adjust their demands downward (price induced conservation). In this case, the fixed demand assumption would seriously bias the results towards high technology and capital-intensive energy supply alternatives. To overcome this deficiency, an effort [12] is under way to combine the Brookhaven model to the Hudson and Jorgenson model [5], so that the effects of changes in the energy supply program can be reflected in the end-use demand behavior. This type of linking the models of different aggregation levels are receiving increasing attention among modelers as a future direction of research (see Ahn [13], Swift [14]).

5.2. The Project Independence Evaluation System (PIES) Model [1]

The Brookhaven models tried to overcome the biased implications of the fixed demand assumption by linking the models to a macroeconomic model. However, the PIES model accomodates directly the price responsiveness of consumers within the model. Thus, the demand side input to the model is not the fixed demand levels, but the estimated market demand functions where the prices of the energy products are included as independent variables. The supply side of the PIES model is similar to that of the BESOM in that it has a linear programming (LP) structure. It combines the ouputs from each of the other components, and solves for market clearing prices, supplies and demands. Step function approximations to resource supply curves are taken from the special-purpose individual supply analyses, and are embedded within the LP structure of the PIES model. This step function approximation cannot be done directly with the econometric demand functions, since there the demand for a particular energy product depends both on its own price (via own-price elasticities) and also on the prices of competing fuels (via cross-price elasticities). Due to cross-price elasticity terms, PIES econometric demand functions cannot in general integrate into a 'social welfare function'. This indicates that conventional optimization techniques cannot be employed to solve for an equilibrium. Alternatively, an iterative procedure is employed. At each iteration, the cross-price elasticities are neglected, but the own-price effects are incorporated through step-function approximations within the LP structure of the PIES model. This augumented LP problem can be solved

by available LP computer packages. At each optimal solution to the augmented LP problem, there are dual variables (long run incremental costs) which serve as inputs to the demand curves for individual fuels. Based on this new supply price information, a new set of own-price demand function (without cross-price elasticities incorporated). The new demand curves are again incorporated within the PIES LP structure through the step-function approximation, a new optimal solution is produced, and this process is repeated until a satisfactory market equilibrium is obtained.

From a computational viewpoint, this PIES iterative approach can be viewed as one of decomposition methods. Ahn [13] has shown that this approach has some theoretical implications which can be applied to many economic problems, and in particular indicated that this approach can be viewed as one of standard approaches in solving for a market equilibrium (not only for the energy sector).

The PIES has proven itself to be a useful practical tool. It was used extensively in the Project Independence Report [15] of the Federal Energy Administration of USA (now the Department of Energy). Unlike standard optimization methods, the PIES permitted policy analysts to evaluate legislative proposals for pricing formulae other than those based on dual variables.

The major limitations of the PIES model arise from the fact that it is a static model of long-run economic equilibrium. It cannot be used to investigate the dynamic interactions among the various elements of the energy sector, or the short-run adjustments necessary for the successful attainment of long-run equilibrium. This type of model cannot specify either the short-run or long-run effects of exogenous shocks to the energy system — such as an Arab oil embargo or the invention of a more efficient method of energy conversion. The analysis of these situations requires a dynamic model that incorporates short-run adjustments and inefficiencies.

5.3. SRI/Gulf/DFI Energy System Model [16]

This model deviates from BESOM in that price sensitive end-use demand curves are incorporated, and that the dynamics of the energy sector through time has been modeled by considering the capacity expansion activities, learning curves, various time lag mechanisms and resource depletion effects. This model was originally developed by the Stanford Research Institute (SRI) and Gulf Oil Corporation to analyze alternative synthetic fuel investments for Gulf. This model is similar to BESOM and PIES in that the energy system components are described and linked within a network starting from various resource extration activities to the specific functional

end-use demands (e.g., space heating, air conditioning) via various conversion processes and transportation/distribution activities. An important and unique feature of this model is the use of long-run supply curves for primary resources (e.g., coal, crude oil and natural gas). These supply curves explain the relationships between the cumulative production and the marginal cost of production. The prices of primary resources are given by the marginal cost from these curves plus an economic rent term (Hotelling [17]) that represents resource scarcity.

End-use demand curves are also estimated as inputs to the model. Unlike the PIES model, demand curves here are based on specific functional end-use (rather than on a specific energy product), and in general cross-price elasticity terms disappear.

The unique market equilibrating algorithm employed in this model allows the decentralized decision making of each component of the network, which is a very useful feature in a large scale modeling of the system. This decentralized nature allows each component to form a separate model which itself might be a LP representation, a set of mathematical relations or a simple equation. In this sense, we can view the SRI/Gulf/DFI model as collection of submodels linked via network flows from 'sources' to 'destinations'.

The basic idea of the network iteration algorithm is to start with initial rough estimates of prices and quantities for all the periods and for all energy products produced or transported by the processes and then successively adjust these prices and quantities until all of the relations embedded in the process models are satisfied. The resulting set of prices and quantities is usually called the equilibrium solution since it is the solution that balances or satisfies all of the relations embedded in the processes. The equilibrium solution, however, will reflect whatever market imperfections and human behavior built into the processes including shortages introduced by price controls and other restriction. This algorithm takes advantage of the inherent network structure of the model in iterating up (end-use demands) and down (resource productions) the network (see Figure 2) computing tentative prices on the upward iteration and quantities on the downward iteration. The upward pass on a given iteration begins with estimates of primary resource prices and computes product prices and end-use energy costs based on a subset of the relations in each process. On the downward iteration quantities are computed using the remaining relations in each process. In iterating up and down the network, the prices or quantities of each process output or input are computed for all time periods before moving to the next output or input. This one way nature of the algorithm is what allows for the decentralized behavior of each component of the system.

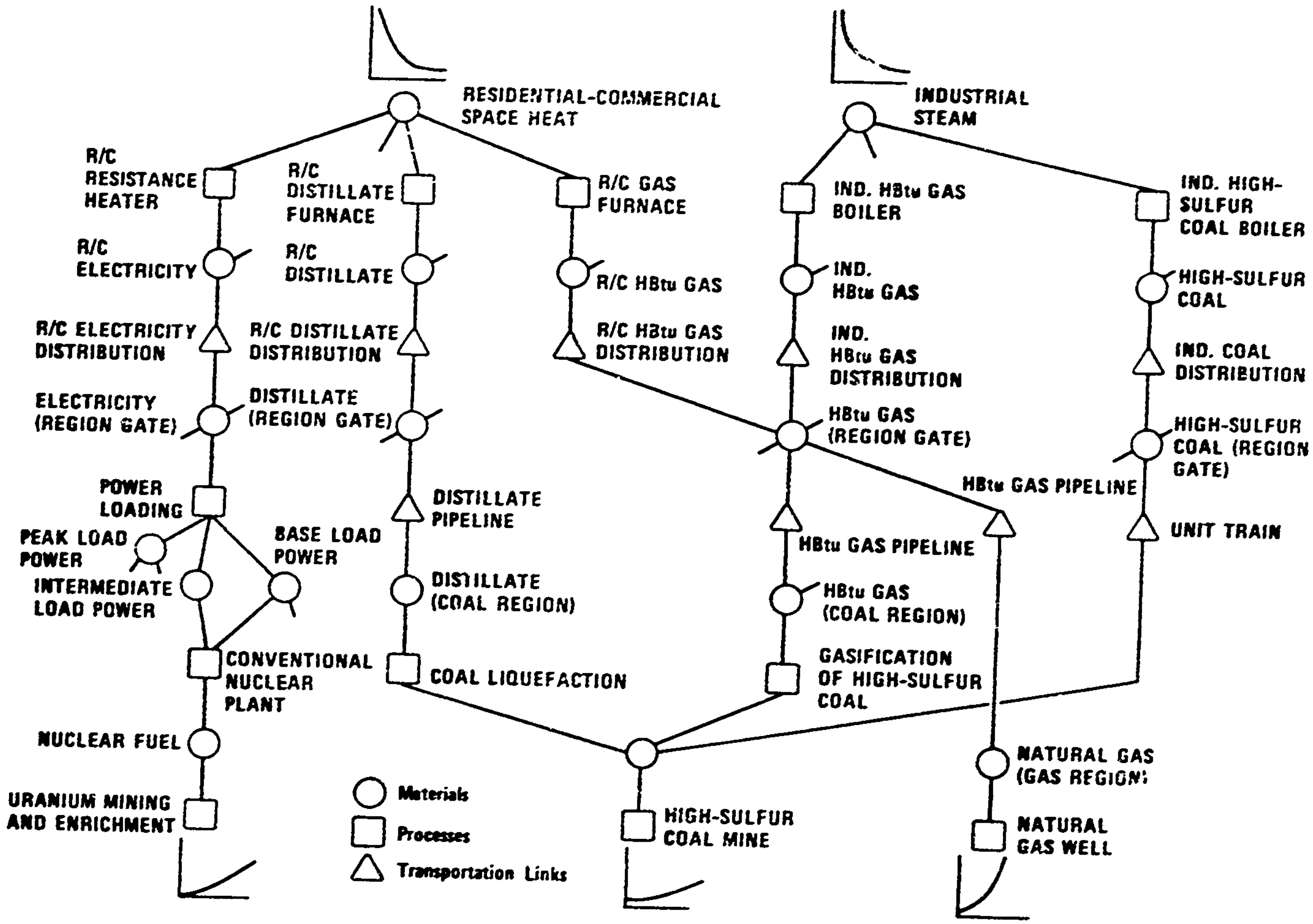


FIGURE 2 EXAMPLE OF ELEMENTS OF THE ENERGY NETWORK

This model has received attention among users, and has been used for analysis of a wide range of corporate and government energy decision problems [18,19]. However, this model is in no way a small-sized one. It has about 100,000 equations and unknowns [16], and the amount of input data is also of a considerable size. It is highly probable in practical applications to make a silly mistake in data preparation making output results incomprehensible. To make one way pass up and down the network, it was necessary to introduce some heuristic modifications which cannot be well supported by relevant economic theories.

5.4. Energy Technology Assessment (ETA) Model

Another model of interest is ETA model developed by Alan Manne [2] of Stanford University. This model's main application was in the nuclear issues -- in particular, nuclear moratorium. Unlike the highly disaggregated specification of end-use demands in the previous model, ETA model has only two energy forms to be finally consumed: electricity and non-electric energy. The supply side of ETA is handled through a conventional linear programming process analysis. Electric energy can be produced by coal-fired power plants, light water reactors, fast breeder reactors and an 'advanced' electric technology (e.g., solar, fusion or an advanced breeder). Non-electric energy (liquids or gases) may be supplied by oil, natural gas, coal- or shale-based synthetic fuels, or hydrogen via electrolysis.

An intertemporal market equilibrium for the energy sector is approximated through a nonlinear optimization algorithm. The objective function here is the sum of consumers' surplus and producers' surplus. The maximization of this surplus over the feasible region leads into the market equilibrium (see Samuelson [20]) under the implicit assumption that the cross-price elasticities satisfy the integrability condition (see Ahn [13]). This was not the case with PIES where the cross-price elasticities were not assumed to be integrable and the optimization techniques cannot be applied.

ETA has been employed primarily in the nuclear power debates. It was one of several models used by the Modeling Resource Group of the Committee on Nuclear and Alternative Energy Systems (CONAES) to estimate the economic value of new energy technologies. ETA was also employed by the Nuclear Energy Policy Study Group (NEPS) to assess the economic costs of foregoing plutonium reprocessing [21] . The NEPS study has been frequently cited as the analytical basis for the Carter Administration's opposition to plutonium-fueled reactors.

This model is relatively well accepted due to its simplicity and reasonable data requirements. But, this could

Energy policy models of this category, in particular, those developed in the United States, have been well documented in the first study [6] of the Energy Modeling Forum of Stanford University. Forthe purpose of completeness, a brief discussion of each model will be included here.

6.1. Hudson-Jorgenson Model — Generalized I-O Approach [5]

The production sector of this model utilizes a nine sector input-output accounting framework with five energy sectors and four nonenergy sectors. Capital and labor are treated as homogeneous quantities along with energy and materials in a production function for each sector. The aggregate energy and material inputs for each sector are further seggregated into separate production functions for the five energy inputs and the four material inputs respectively. Hence, there is a hierarchical structure of 27 production functions which combine to provide implicitly the nine production functions in terms of the nine products.

Interfuel substitution across the five macroenergy sectors and factor substitution across labor, capital, material, and energy are modeled explicitly using econometric relationships. Intrafactor perfect substitutability among types of capital and among types of labor is implicit in the assumption of homogeneous capital and labor.

Interactions of variables are myopic. Thus, prices and quantities determined in the production process depend only on the current period. The link over time, beyond exogenous trends, is found in the transfer of aggregate capital services. To the extent that capital services adjust gradually over time, the reponse of the production sector is gradual. However, the reponse of the production sector to price changes is instantaneous.

The nine sector accounting of the production sector is repeated in the final demand categories, which are further separated into consumption,investment, government expenditures, exports, and imports. However, in computing the tradeoffs between consumption and investment, or labor and leisure, the nine sectors are aggregated to one, with rules to insure consistency of values and prices. The disaggregation of quantities is through fixed shares for iivestment, government, imports, and exports. A series of behavioral relations with nonzero price elasticities is used for the disaggregation of consumption into the nine sectors.

Substitution using econometric relationships occurs at the aggregate level between consumption and investment, and between labor and leisure. Given the aggregate values, there is no substitution across the nine sectors for investment goods, government expenditures, exports, or imports. For

consumption, however, some substitution across energy and materials is included via econometrically estimated constant elasticity price effects. Also, an econometric representation of interfuel substitution is included in the model.

The behavioral relations governing the tradeoffs between aggregate consumption and investment are based on a formulation implying optimization over time. Through simplifying assumptions, this is implemented as a series of myopic calculations. Thus, the determination of consumption also is myopic and depends only on the corresponding prices and quantities in the current period.

The producers and consumers interact in markets where prices and quantities are in equilibrium. Producers demand capital services and labor which are obtained from consumers. Conversely, consumers demand consumption goods and leisure. Hence, capital formation and labor participation are determined endogenously. In each market, equilibrium is determined through the behavioral equations when the supply-demand prices and quantities for all transactions are equal.

The balance is determined sequentially in each period with the available homogeneous capital services operating as the dynamic link. Hence, separately for each period, the model determines a general market equilibrium in all markets.

6.2. Wharton Annual Energy Model [23]

The Wharton model is a highly disaggregated system evolving from the large Wharton EFA annual and interindustry system.

The model incorporates 59 industrial output sectors of which eight are energy producing sectors and the remainder produce various nonenergy goods and services. Separate production functions are estimated for each sector using a two level hierarchy in which, for each sector, Cobb-Douglas production function is used to determine value added from inputs of labor and capital services, and a constant elasticity of substitution, multivariable production function to determine aggregate level of intermediate inputs as a function of the vector of intermediate inputs, and a final production function assuming perfect complementarity between value added and aggregate intermediate inputs to determine the sectoral output.

Interfuel substitution across energy sectors, factor substitution across labor and capital, and substitution across intermediate inputs is modeled through econometric relationships using the mathematical structure outlined above. Intrafactor perfect substitutability is implicit in the assumption of homogeneous capital and labor. No direct substitution between material and factor inputs is included.

By changing the mix of sector outputs, however, indirect substitution between materials and other factors is included.

The variable input-out coefficients are determined in the long run by the prices of all factor inputs. However, the adjustment to the long run values is not instantaneous. As implemented, in any period the coefficients depend on current and previous period prices and the rate of adaptation is determined by a separate lag parameter.

Final demand is decomposed into consumption (14 categories), investment (32 categories), inventories, trade (14 categories) and government (6 categories). Behavioral equations for each of these categories are included in the macroeconomic model. Each category then is disaggregated in turn into the 59 sectors of the interindustry classification through the application of fixed shares.

Substitution using econometric relationships occurs in all final demand categories at an appropriate level of aggregation noted above. These substitutions are a major source of potential variability in total energy use across major categories by final demand. Due to the extensive detail, the effect of substitution assumptions implicit in the level of aggregation is minimal.

The consumption equations operate on lagged prices and quantities which replicate a gradual adjustment ot long run equilibrium. There is a separate parameter controlling the speed of adustment for each final demand category weighted by the price and quantity differences in the previous period. No direct consideration of future prices is included.

Long run equilibrium at full employment is a target for the model subject to the constraints implied by the dynamics of the short and long run adjustments. There is full short run equilibrium in the product markets in terms of prices and quantities. However, there is some uncertainty as to what extent the interindustry demands for capital and labor inputs are balanced with the prices and supply determined in the macroeconomic model of consumption-investment or employment-unemployment.

The model does not consider future prices as relevant to the decisions in any period. Therefore, the solution implementation is sequential, computing a short run equilibrium in each period. Due to the consideration of past prices and quantities, neither full employment nor long run equilibrium is achieved in any period. Rather, these serve as targets which the model approaches gradually.

6.3. DRI-Brookhaven Model [24]

The implementation of this linked system is achieved through information tranfers for three target years of 1985, 1990, and 2000 between the Hudson-Jorgenson (DRI) model,

and the combined Input-Output-BESOM model of the Brookhaven National Laboratory. The Hudson-Jorgenson model described earlier is used as an intertemporal integrating device with the static I/O-BESOM model providing energy technology detail for the three target years. The information of aggregate energy demands for three target years at five sector detail is transmitted from the Hudson-Jorgenson model to the I/O-BESOM model. The detailed I/O-BESOM model in turn determines the relative prices, the fuel mix, and the capital requirements for energy taking into account the availability of new energy technologies and interfuel substitution is achieved through an eight order disaggregation of the end use categories (such as space heat, process heat, petrochemical feedstocks, motive power, etc.). While the I/O-BESOM model's computer implementation for these target years is independent, separate numerical checks are made to assure intertemporal consistency of energy conversion and end use capacities.

Since the Hudson-Jorgenson subsystem possesses the characteristics of their model described earlier in this paper, they will not be repeated here. We briefly note, however, that the industrial sectors are disaggregated into nine sectors with five sectors for energy production. The labor and capital are treated as homogeneous quantities and the main dynamic link is provided through capital services. The market equilibrium is myopically determined for each period through behavioral equations for production and consumption.

6.4. ETA-MACRO Model [25]

ETA-MACRO (Figure 3) is an expanded version of Alan Manne's Energy Technology Assessment (ETA) model. It adds a very simple representation of the interaction between the energy system and the rest of the economy to ETA's process representation of the energy supply system.

Oil and gas, aggregated together, coal and uranium are explicitly represented in the model. Other fuels, hydro, geothermal, etc. are specified exogenously. Oil refining is not considered in the model; electric conversion is given an explicit process representation, and coal synthetics are represented in terms of their market price. Coal and nuclear fuels are the two possible choices for baseload electricity generation; the load structure of electricity demands is not considered. The model is dynamic and assumes perfect foresight on the part of energy producers. A thirty year plant life is assumed.

The economy is assumed to produce a single aggregate form of output from capital, labor and energy inputs. A macroeconomic growth model provides for substitution between capital, labor and energy inputs. The energy inputs are disaggregated into electricity (measured at the busbar)

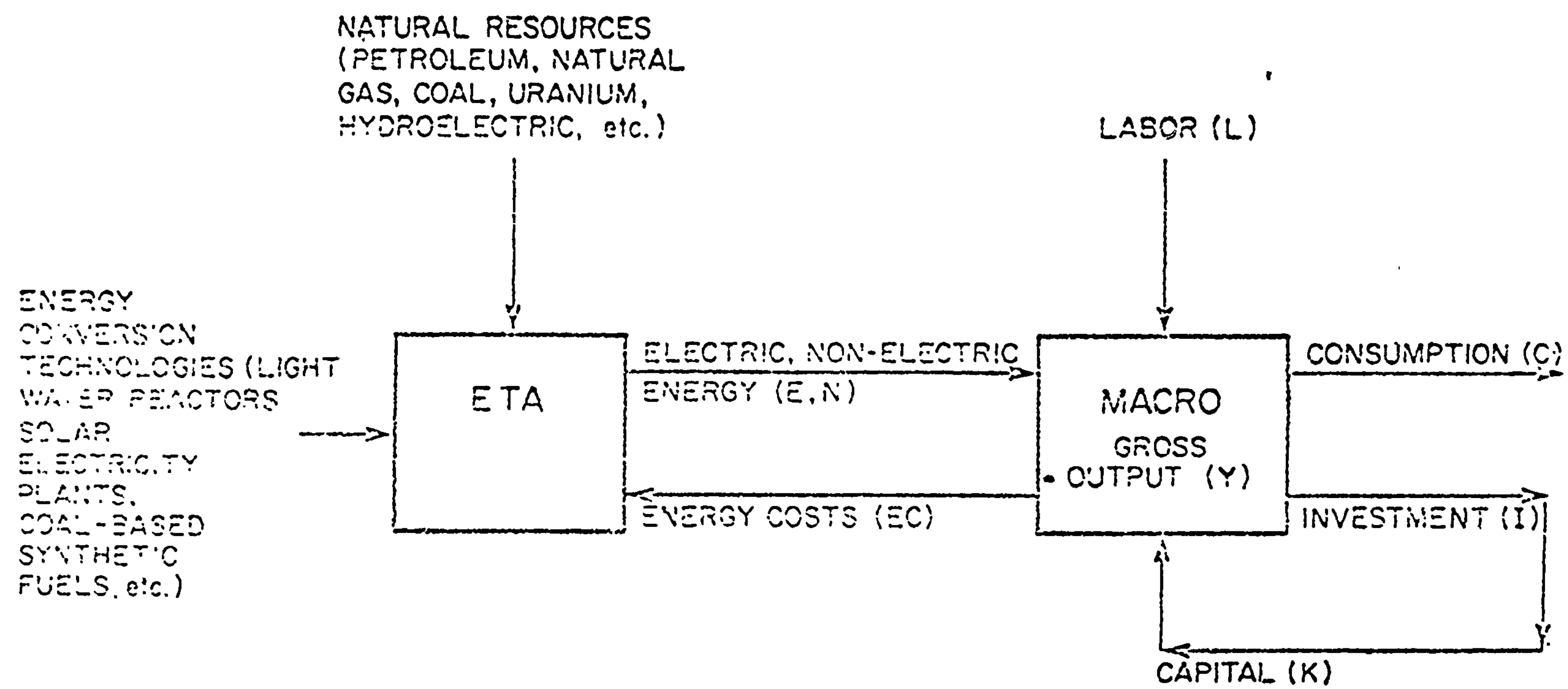


FIGURE 3 AN OVERVIEW OF ETA-MACRO

and nonelectric energy (measured at the refinery gate). A Cobb-Douglas production function provides for substitution between the two energy forms. The demand side of the model is dynamic and assumes perfect foresight on the part of energy demander. For example, savings and investment decisions are modeled so that consumers will receive equal benefits from an additional dollar's worth of current consumption and a dollar's worth of investment. Society's capital stock is assumed to be retired at a constant rate.

As was the case with ETA model, this model is welcomed for its structural simplicity and reasonable data requirements. This model's results, however, cannot be used in a detailed investment planning processes due to the lack of relevant detailed description of the technologies. This has been successfully used in evaluating the policy questions which have long-range implications upon the economy such as nuclear moratorium.

Methodologically, this model serves as a good example of model linking exercises (here the macroeconomic growth model and the process analysis energy sector model) which is one of current research trends in the energy modeling community.

7. Linking Energy Models

We have examined briefly scopes, applications, methodology and major characteristics of various energy models used for analysis of various energy policies. This review is not exhaustive, but rather is intended to illustrate the structure of recent and current efforts among energy modeling community. This review does not serve as a catalog of representative models. Rather, the discussion was focused on important issues and properties of energy models, around which appropriate models are identified as examples satisfying relevant properties under discussion.

Modeling efforts in energy areas have been classified mainly in terms of the market sizes associated with policy questions: single fuel market models, energy system models and energy-economy integrated models. Typically, those for a single fuel can be modeled at a considerably detailed level including the technical and engineering details of supply technologies and consumption technologies. Thus, the output results from these models may be used directly as implementable guidelines and references.

On the other hand, models which consider the energy sector as a whole and the rest of the economy within a single framework cannot model the details of a particular technology due to the size limitation (in terms of available computation capability or data requirements). Outputs from these models typically serve as a general guideline or long-term suggestions

which might not be used in practical situations without supplemental studies.

One of current research interests in energy system modeling is the modular approach. This approach tries to combine separate models for integrated analysis. This permits the use of a variety of expertise, efficient manipulation of data, gradual replacement of old components by new models, and simplified through the isolation of the relevant components for particular planning decisions. The practice has proceeded, however, without analysis of the requirements for model interface. Common techniques, independently developed, have been isolated and subjected to a comparative analysis. A general framework has been developed for defining a combined modeling system in terms of a hierachy or generalized network of models. The requirements for the description of a competitive equilibrium have been derived and the extensions to the analysis of other systems have been indicated. These results are very beneral and can be applied to the analysis of the disaggregated systems of an individual utility, e.g., the analysis of different customer classes, or the aggregated models of national energy supply and demand.

Hogan [26] has pointed out the following areas for research extension in model linking efforts:

(1) the development of improved process models,
(2) design of improved approximation techniques (model interface mechanisms),
(3) the analysis of the theory of descriptive models,
(4) the establishment of criteria for optimal model disaggregation,
(5) the development of improved software, and
(6) the investigation of several facets of the computational problems inherent in the use of modular systems.

Approximation methods are popular for constructing model interfaces, whether through the use of pseudo data, extreme point representations, outer linearization, or the construction of simple models to represent a more complicated model. For example, in the PIES model, the demand side submodel is approximated as a constant elasticity demand curves in order to be linked to the supply sector model at each iteration of the computation. Griffin [27] has suggested to use a detailed model to generate a set of data (called pseudo data) which is to be fitted by some simple statistical model. The error properties of colinear experiments that occupy so much attention in econometric estimation can be overcome in the

experimental design for generating pseudo data.

The movement to combined energy models has progressed simultaneously with the recognition that decoupled iterative procedures can work as well as a more structured optimization or equation solving approaches. The experiance of the PIES model and the SRI/Gulf/DFI system are particularly relevant here. There have been substantial theoretical advances in this area recently, including the theoretical studies of Ahn [13] and Thrasher [28] in characterizing the convergence properties of the PIES algorithm. This work has demonstrated a close link between the price-quantity iterative procedures and approximate Jacobi methods for the solution os systems of equations. The generalization of these ideas to a network of models and a price quantity iteration indicates that it is possible to incorporate approximation and relaxation procedures for the acceleration of convergence, using a very simple process of decentralized iteration, balancing locally approximate supply and demand curves.

The author is experimenting one of decoupled iterative approach in solving an integrated energy/economy model to examine the effects of the availability of crude oil and the increasing prices of energy products upon the economic growth of a nation. The economy of the nation is described by one of popular optimal economic growth models employing a national production function whose independent variables include capital, labor and energy (e.g., electricity and non-electric energy). The energy sector is represented by a dynamic process analysis model which tries to minimize the discounted sum of capital and operating costs to meet specified demands of final energy products. By linking these two models, it becomes possible to analyze the effects of energy sector behavior upon the rest of the economy.

The model formulation itself is not new at all. For example, the Manne's ETA-MACRO [25] has formulated a model in which the energy sector is described in terms of process analysis approach, and the rest of the economy is represented by an economic growth model. This model, however, tries to solve for solutions within a single framework employing an optimization method rather than a decoupled iterative approach.

It is highly desired that more research efforts are given to the establishments of theoretical framework and computational methods in combining energy models. This will one of research efforts which will occupy energy system modelers' minds for some time to come.

REFERNCES:

1. PIES Documentation, "The Integrating Model of the Project Independence Evaluation System," vol. 1, FEA/N-76/411, Washington, D.C., USA, September 1976.

2. Manne, A.S., "ETA: A Model for Energy Technology Assessment," Bell Journal of Economics, vol. 7, no. 2, Autumn 1976.

3. Cherniavsky, E.A., "Brookhaven Energy System Optimization Model," Brookhaven National Lab. Topical Report No. BNL 19569, Associated Universities Inc., Upton, NY, USA 1974.

4. Christensen, L.R., D.W. Jorgenson, and L.J. Lau, "Transcendental Logarithmic Production Frontiers," Review of Economics and Statistics

5. Hudson, E.A., and D.W. Jorgenson, "US Energy Policy and Economic Growth, 1975-2000," Bell Journal of Economics and Management Science, vol. 5, no. 2.

6. Energy Modeling Forum, Energy and the Economy, EMF Report 1, volumes 1 and 2, September 1977.

7. Kennedy, M.,"An Economic Model of the World Oil Market," Bell Journal of Economics and Management Science, vol. 5, no. 2, 1974.

8. Baughman, M.L., and P.L. Joskow,"The Future of the US Nuclear Energy Industry," Bell Journal of Economics, vol. 7, no. 1, Spring 1976.

9. Marcuse, W., Bodin, L., "A Dynamic Time Dependent Model for the Analysis of Alternative Energy Policies," Brookhaven National Laboratory, Upton, NY, USA.

10. Behling, D.J., Jr., and Dennegy, G., "Estimation of Final Demands and Current Input-Output Coefficients to 1985," Report BNL 20369, Brookhaven National Laboratory, Upton, NY, August 1975.

11. Behling, D.J., Jr., R.C. Dullien, E.A. Hudson,"The Relationship of Energy Growth to Economic Growth under Alternative Energy Policies," Report BNL 50500, Brookhaven National Laboratory, Upton, NY, 1976.

12. Hoffman, K.C. and D.W. Jorgenson, "Economic and Technological Models for Evaluation of Energy Policy," Bell Journal of Economics, vol. 8, no. 2, 1977.

13. Ahn, Byong-hun, Computation of Market Equilibria for Policy Analysis: The Project Independence Evaluation System (PIES) Approach, Garland Publishing Inc., New York, N.Y., 1979.

14. Swift, M.,"On linking a Brookhaven Model to a Macroeconomic Model," PhD Dissertation, Department of Engineering-Economic Systems, Stanford University, 1978.

15. National Energy Outlook, Federal Energy Administration, Washington, D.C., USA, February 1976.

16. Cazalet, E., "Generalized Equilibrium Modeling: the Methodology for the SRI-Gulf Energy Model," Decision Focus Inc., Palo Alto, California, May 1977.

17. Hotelling, H., "The Economics of Exhaustible Resources," Journal of Political Economy, April 1931, vol. 39.

18. Synfuels Interagency Taskforce, "Recommendations for a Synthetic Fuels Commercialization Program," Washington, D.C.: US Government Printing Office, Stock Number 041-001-00111-3, November 1975.

19. Stanford Research Inst.,"A Western Regional Energy Development Study: Economics," volumes 1 and 2, November 1976.

20. Samuelson, P.A., Foundation of Economic Analysis, Cambridge, Massachusettes, Harvard University Press, 1947.

21. Report of the Nuclear Energy Policy Study (NEPS) Group, Nuclear Power Issues and Choices, Ballinger Publishing Company, Cambridge, Massachusetts, 1977.

22. Hogan, W.W., and Manne, A.S., "Energy-Economy Interactions: The Fable of the Elephant and the Rabbit?", Working Paper EMF 1.3, Energy Modeling Forum, Stanford University, Stanford, California, USA, July 1977.

23. Klein, L.R., and Finan, W.F., "The Structure of the Wharton Annual Energy Model," (paper prepared for the Energy Modeling Forum Meeting, Washington, D.C., October 1976), Wharton EFA, Philadelphia, Pa., USA, 1976.

24. Jorgenson, D.W., et al, "Integration of the Brookhaven National Laboratory Engineering-Oriented Energy Model and the Data Resources, Incorporated Econometric-Oriented Total Industrial Model," Final Report to the Brookhaven National Laboratory, by Data Resources, Inc., Cambridge, Mass.

25. Manne, A.S., "ETA-MACRO: A Model of Energy-Economy Interactions," Report EPRI EA-592, Electric Power Research Institute, Palo Alto, CA, December 1977.

26. Hogan, W.W.,"On Linking Energy Models," mimeographed, Energy Modeling Forum, 1978

27. Griffin, J.M.,"The Process Analysis Alternative to Statistical Cost Functions: An Application to Petroleum Refining," AER, vol. 62, March 1972.

28. Thrasher, R.,"Notes on Convergence of PIES-like Iterative Processes," PhD Dissertation, New Mexico State University, 1979.

THE CAPACITATED P-MEDIAN MODEL FOR THE LOCATION OF FACILITIES IN THE PUBLIC SECTOR.

Prof. Dr. . KAUFMAN, Vrije Universiteit Brussel,
Pleinlaan 2, B-1050 Brussels, Belgium.

Prof. Dr. F. BROECKX, Universiteit Antwerpen (RUCA)
Middelheimlaan 1, B-2020 Antwerp, Belgium.

ABSTRACT. The problem of locating p facilities in the public sector, with as objective the minimization of the total distance travelled by the users has been extensively studied and many algorithms have been published.

In this paper we focus on an extension of this problem in which the sizes of eventual facilities are limited by capacity constraints. It is clear that such constraints make the model more realistic and applicable to real public sector location problems, especially when the model is used for the extension of an existing network of facilities.

The problem is formulated as a mixed integer program which can also be used for comparing alternative solutions proposed by planners. In the paper several applications are presented such as hospital implantation, location of schools, fire stations, etc.

In order to find good or optimal solutions an exact and an heuristic algorithm will be presented. The exact algorithm is based upon the Benders decomposition technique.

The paper concludes with a discussion of a computer program and computational experience.

1. INTRODUCTION

In a previous paper [1] we proposed an exact algorithm for the p-median problem based upon the Benders' decomposition approach [2]. In this problem a given number p of facilities must be located among a given set of m candidate sites situated on a graph or a network. There are important differences between the treatments of location problems in the public and private sectors. These have been extensively discussed by REVELLE, MARKS and LIEBMAN [3].

In the private sector where the location problems concern plants, warehouses, bank outlets, etc. in a network of more or less concentrated users, one usually has well defined objectives usually expressed in monetary terms (maximizing profits or minimizing costs).

In the public sector the problems are for example the implantation of hospitals or firestations in a city, ambulance dispatch centers in a region, schools, post-offices or other public facilities. The criteria used here are related to surrogate measures of social welfare. The utility functions are mostly expressed in terms of physical distances to be covered, or time spent by the users or so-called response times, covering both distance and time. Minimizing the total distance covered by the users leads to the p-median problem. Minimizing the maximal distance to be covered, results in the p-center problem.

The basic theoretical and practical ideas on the location of centers and medians were introduced by HAKIMI [4]. Since then several methods and algorithms, both exact and approached have been proposed for both problems; see for example the papers by GARFINKEL, NEEBE and RAO [5], BAKER [6] JÄRVINEN, RAJALA and SINERVO [7].

An important situation in which models of this type can be used, occurs when an extension of an allready existing system of facilities is being considered. In this case and also very often with new facilities it is necessary to consider capacity limitations of the facilities.

A similar development can be seen in private sector models. After the first model for the simple plant location problem, proposed by BALINSKI [8] further developments have led to more sophisticated models in which capacities were considered. Important results were obtained by SA [9], DAVIS and RAY [10], and GEOFFRION and GRAVES [11].

In this paper we consider the following generalization of the p-median problem: "Given a set of locations where

a facility can be located and a set of population centers to be served by the facilities, given capacities on eventual facilities and the number of users in each population center, given the number p of facilities to be located; determine the locations and sizes of the facilities to be built, in order to minimize the total distance travelled by the users".

A similar type of constraint as the capacity constraint could be considered for expressing cost limitations for facilities in a given region. Another possible generalization concerns the number p of facilities. Parametrizing this number can be usefull for political decision making.

In the next section we consider a mixed-integer programming model for solving this problem.

2. MATHEMATICAL MODEL

The problem described in the previous section can be formulated as a mixed integer programming problem. This requires the following notations:

- the number of sites in which a facility can be built is called m; a subscript i ($i=1,2,\ldots,m$) will denote such a site.
- the number of users, or population centers, is called n; a subscript j ($j=1,2,\ldots,n$) will denote a population center.
- p_j is defined as the number of users in a population center j.
- a_{ij} is the average distance between a site i and a population center j.
- $d_{ij}=a_{ij}p_j$ is the total distance travelled by the users of population center j if they all use a facility located in i.
- c_i is the capacity of a facility at site i.
- p is the number of facilities to be located.

The problem also requires the following variables:

- t_{ij} is the fraction of users from population center j going to facility i; it is clear that the distance travelled by the users is given by $a_{ij}p_jt_{ij}=d_{ij}t_{ij}$.
- y_i is a 0-1 variable which describes the selection of site i for establishing a facility: if a facility is established at site i, y_i is equal to one, and it is equal to zero if no facility is located in i.

The objective of the problem is to minimize the total distance travelled by the users. As the distance travelled by the users from center j going to facility i is equal to $d_{ij}\, t_{ij}$ the total distance travelled by the users from center j is:

$$\sum_{i=1}^{m} d_{ij} t_{ij} \tag{1}$$

The total distance travelled by all users is then equal to:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} t_{ij} \tag{2}$$

and the objective will be to minimize

$$\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} t_{ij} \tag{3}$$

The variables t_{ij} and y_i must satisfy a set of constraints which will now be discussed:

- At first it must be expressed that each of the p_j users from population center j must be assigned to a facility. As t_{ij} is the fraction of these users assigned to facility i the following constraint must hold:

$$\sum_{i=1}^{m} t_{ij}=1 \qquad , \; j=1,2,\ldots,n \tag{4}$$

- An important new constraint results from the introductior of capacity limitation on the facilities:

$$\sum_{j=1}^{n} p_j t_{ij} \leq c_i y_i \qquad , \; i=1,2,\ldots,m \tag{5}$$

These constraints have a double function: if there is a facility at site i and therefore y_i is equal to one the constraint becomes:

$$\sum_{j=1}^{n} p_j t_{ij} \leq c_i \tag{6}$$

It expresses that the total number of users of facility i (per time period) does not exceed c_i, the capacity of facility i . If no facility is located at site i the constraint becomes:

$$\sum_{j=1}^{n} p_j t_{ij} \leq 0 \tag{7}$$

which can only be satisfied if for that specific i all t_{ij}

are zero, which simply means that if there is no facility at site i no one can use such a facility.

- The fact that exactly p facilities are to be located is expressed as:

$$\sum_{i=1}^{m} y_i = p \qquad (8)$$

- Finally by the nature of the variables considered the following constraints are obvious

$$y_i \in \{0,1\} \quad , \quad i=1,2,\dots,m \qquad (9)$$

$$t_{ij} \geq 0 \quad , \quad i=1,\dots,m; \quad j=1,\dots,n \qquad (10)$$

The model represented by the objective (3) and constraints (4), (5), (8), (9) and (10) can then be written as:

$$\text{minimize} \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} t_{ij} \qquad (11)$$

subject to

$$\sum_{i=1}^{m} t_{ij} = 1 \quad , \quad j=1,\dots,n \qquad (12)$$

$$\sum_{j=1}^{n} p_j t_{ij} \leq c_i y_i \quad , \quad i=1,\dots,m \qquad (13)$$

$$\sum_{i=1}^{m} y_i = p \qquad (14)$$

$$y_i \in \{0,1\}, \quad i=1,\dots,m \qquad (15)$$

$$t_{ij} \geq 0 \quad , \quad i=1,\dots,m; \quad j=1,\dots,n \qquad (16)$$

This problem will be referred to as the capacitated p-median problem (C P M P).

It is a generalization of the p-median problem on a graph which was stated by HAKIMI [4] as follows:
"Given a graph with values associated to both the edges and the nodes, select p nodes on the graph for which the sum of the products of the distances of each of the nodes to the nearest of the p nodes by the value associated with that node is minimum".

Another remark which can be made concerns the variables t_{ij}. Taking the t_{ij} to be zero or one would impose that the users from one population center all go to the same facility, which does not seem all to realistic.

Finally it can be observed that more general economic functions than the total distance travelled can be conside-

red. Also replacing the a_{ij} by any costs proportional to the number of users makes it possible to consider any linear cost structure. Using piecewize linearization makes it possible to consider general convex cost functions.

It is clear that the problem grows very quickly, because for locating p facilities among m candidate sites, with n population centers the number of variables (y_i, t_{ij}) is m(n+1) and there are m+n+1 constraints.

3. AN EXACT ALGORITHM USING BENDERS DECOMPOSITION

The model described by equations (11)-(16) is of the mixed integer type. It can be observed that when the y variables have been fixed the constraints can be transformed into transportation problem type constraints. Indeed consider

$$x_{ij}=p_j t_{ij} \tag{17}$$

or

$$t_{ij}=x_{ij}/p_j \ , \tag{18}$$

the first constraints (12) give:

$$\sum_{i=1}^{m} t_{ij}=\sum_{i=1}^{m} x_{ij}/p_j=1; \quad j=1,\ldots,n \tag{19}$$

or

$$\sum_{i=1}^{m} x_{ij}=p_j \ , \quad j=1,\ldots,n \tag{20}$$

and the second constraints (13) become:

$$\sum_{j=1}^{n} x_{ij}\leq c_i y_i; \quad i=1,\ldots,m \tag{21}$$

As the coefficients d_{ij} are equal to $a_{ij}p_j$ the model can now be written as

$$\text{minimize} \ \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}x_{ij} \tag{22}$$

subject to

$$\sum_{i=1}^{m} x_{ij}=p_j \ , \quad j=1,\ldots,n \tag{23}$$

$$\sum_{j=1}^{n} x_{ij}\leq c_i y_i \ , \quad i=1,\ldots,m \tag{24}$$

$$\sum_{i=1}^{m} y_i=p \ , \tag{25}$$

$$y_i \in \{0,1\} \quad , \qquad i=1,\dots,m \tag{26}$$

$$x_{ij} \geq 0 \quad , \qquad i=1,\dots,m; \quad j=1,\dots,n \tag{27}$$

When the y_i variables are fixed at 0 or 1 and the p_j play the part of demands and the c_i that of supplies one obtains a classical transportation problem.

To solve this problem by taking advantage of the special structure of constraints (23) and (24) two approaches seem indicated. The first consists in applying an implicit enumeration procedure by considering all possible values of the y-variables. The second approach which will be described here in detail consists in applying the Benders decomposition algorithm to the mixed integer problem (22)-(27).

The algorithm consists in the following steps:

<u>Step 1</u>: initialization.

- Find a first feasible solution for the y variables and call it $y^H=(y_1^H, y_2^H,\dots,y_m^H)$, with $H=1$

  The first solution vector can either be obtained using an heuristic such as the one described in the next section or simply taking p facilities with sufficient capacity to be able to satisfy the demand from all population centers.
- Set $z^{opt}=\infty$. z^{opt} will denote the value of the best solution found so far. This value is an upper bound on the value of the optimal solution.
- Set $\underline{z}=0$: $\underline{z}$ is a lower bound on the value of the optimal solution.
- Fix a tolerance parameter T. It is considered sufficient to find a solution within Γ of the optimal one.
- Go to step 3.

<u>Step 2</u>: 0-1 master problem.

- In the 0-1 master problem constraints containing 0-1 variables only, in our problem only constraint (25), are considered together with all constraints generated by step 3. These last constraints are called Benders' cuts. This problem can be written as:

$$\text{maximize } y_0 \tag{28}$$

  subject to

$$y_0 \geq \sum_{j=1}^{n} u_j^h p_j - \sum_{i=1}^{m} v_i^h y_i; \qquad h=1,\dots,H \tag{29}$$

$$\sum_{i=1}^{m} y_i = p \tag{30}$$

$$y_i \in \{0,1\} \quad , \quad i=1,\ldots,m \tag{31}$$

where the values u_j^H and v_i^H of the last Benders cut (for h=H) are given by the optimal dual variables found the previous time step 3 is carried out.

- It was shown by BENDERS [2] that the value of the optimal solution of this problem is a lower bound on the value of the optimal solution of the original mixed integer problem. This value will be called $\underline{z}$.
- Increase H by 1 and call the optimal solution of problem (28)-(31):

  $$y^H=(y_1^H,y_2^H,\ldots,y_m^H)$$

  To solve problem (28)-(31) many general purpose algorithms can be used. At present a specific implicit enumeration algorithm is being developed by VAN OUDHEUSDEN and PLASTRIA [12].
- Then go to step 3.

<u>Step 3</u>: continuous problem.

- Solve the transportation problem obtained by fixing the y_i variables to the feasible values y_i^H found in step 1 or 2. This problem is given by

  $$\text{minimize} \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}x_{ij} \tag{32}$$

  subject to

  $$\sum_{i=1}^{m} x_{ij}=p_j \quad , \quad j=1,\ldots,n \tag{33}$$

  $$\sum_{j=1}^{n} x_{ij}\leq c_i y_i^H, \quad i=1,\ldots,m \tag{34}$$

  $$x_{ij}\geq 0 \quad , \quad i=1,\ldots,m \ , \ j=1,\ldots,n \tag{35}$$

  The value of the objective function (32) in its optimal solution, $\bar{z}$, is an upper bound of the value of the optimal solution of the original problem.
- If

  $$\bar{z}<z^{opt} \tag{36}$$

  replace z^{opt} by $\bar{z}$.
  In this case a better solution has been found.
- If

  $$z^{opt}<\underline{z}+T \tag{37}$$

  an optimal solution has been found within a tolerance of T, and the algorithm is terminated.
- Otherwise the optimal dual variables of problem (32)-(35) must be computed. This dual problem can be written as:

$$\text{maximize} \quad \sum_{j=1}^{n} p_j u_j - \sum_{i=1}^{m} c_i y_i^H v_i \tag{38}$$

subject to

$$u_j - v_i \leq d_{ij} \quad , \quad i=1,\ldots,m \; ; \; j=1,\ldots,n \tag{39}$$

$$v_i \geq 0 \quad , \qquad i=1,\ldots,m \tag{40}$$

- Call the optimal dual variables of iteration H, u_j^H and v_i^H
- Go to step 2.

It can be observed that most algorithms for the transportation problem provide immediately optimal dual variables.

4. AN HEURISTIC ALGORITHM

It is expected that the exact algorithm will only be able to provide optimal solutions for small to medium sized problems and therefore it is necessary to use an heuristic for large problems. Such an heuristic is also useful for providing a good initial solution when using the exact algorithm. In this section such an heuristic will be described.

The first solution vector $\mathbf{y}^0$ is obtained by considering all facilities open. For this solution $y_i^0=1$ for all i.

Subsequently at each iteration one facility is closed untill exactly p facilities remain open. The selection of a facility to be closed is done among a subset of facilities satisfying the following constraint: among the remaining open facilities the p facilities with the largest capacities must be large enough to satisfy the demand from all population centers.

To meet this requirement the total demand must first be calculated by:

$$D = \sum_{j=1}^{n} d_j \tag{41}$$

A facility will then be considered for closure if there exists a subset K of p facilities which remain open such that

$$\sum_{i \in K} c_i \geq D \tag{42}$$

At each iteration a facility i is then closed for which the total cost shows the smallest increase. The calculation of the increase requires the following definitions:

$$d_{mj} = \min_{i | y_i^h = 1} d_{ij} \tag{43}$$

and $$d_{kj} = \min_{i \mid y_i^h = 1 \text{ and } k \neq m} d_{ij} \tag{44}$$

d_{mj} and d_{kj} are the smallest and second smallest distances of population center j to an open facility.

The total cost for the present iteration is given by

$$z^h = \sum_{j=1}^{n} d_{mj} \tag{45}$$

If a facility i is being considered for closure the increase in cost $r_j(i)$, for a customer j, is then computed by

1. If $d_{ij} > d_{mj}$ then (46)

$$r_j(i) = 0$$

2. If $d_{ij} = d_{mj}$ then (47)

$$r_j(i) = d_{kj} - d_{mj}$$

Indeed if facility i would have served customer j, this customer will now use k, the second closest facility.

The total increase in costs is given by

$$R(i) = \sum_{j=1}^{n} r_j(i) \tag{48}$$

At each iteration the facility i will be closed for which R(i) is minimum. Iterations will continue until p facilities remain open.

After the final iteration an algorithm for the transportation problem will be used to find the local optimal solution and its value. The implementation of this algorithm will be discussed in the next section.

5. COMPUTATIONAL EXPERIENCE

Both the exact and heuristic algorithms for solving the CPMP are being programmed in FORTRAN for a CDC 6500 computer

For the master problem in the Benders decomposition approach the use of a general purpose 0-1 algorithm has proven quite disappointing. The resolution of a large number of 0-1 programmes of increasing sizes (at each iteration the number of constraints is increased by one) takes a prohibitively large computation time even for small problems. Therefore the general 0-1 algorithm is being replaced by a specialized algorithm due to VAN OUDHEUSDEN and PLASTRIA [12] and which is based upon the analogy between the master problem and a specific problem of game theory. Results of this attempt are forthcoming.

The heuristic algorithm has been programmed in FORTRAN and two sets of experiments have been carried out.

In the first set a region is considered in which 5 facilities are to be located among a number of possible sites which varies between 20 and 50 with steps of 10. The set of customers is taken as identical with the set of possible sites. This set is randomly generated on a plane. The coefficients in the model are also randomly generated from a rectangular distribution with the following characteristics:

p_j: between 5 and 15

c_i: between 100 and 300

The times necessary for solving the problems as well as the total distances travelled are given in table 1.

Problem NO.	NO.of possible sites	Computation time in sec. CPU	Total distance travelled
1	20	5.09	4785
2	30	19.85	7339
3	40	30.02	12952
4	50	82.04	12655

\- Table 1 -

It is clear that a complete analysis of the heuristic should include a comparison of its solution with the optimal one. This together with improvements in the heuristic is being carried out at present.

In the second set of experiments several problems with the same coefficients are solved with the number of facilities to be selected varying between 2 and 10. The coefficients are generated in the same way as for the first set. The computation times and costs are given in table 2. Such

a table is usefull for the decision makers for a study of the trade-off between the costs for establishing a new facility and the total distance travelled by the users.

Problem NO.	NO. of facilities to be located	computation time in sec. CPU	Total distance travelled
1	2	0.76	6240
2	3	2.31	6015
3	4	4.33	4836
4	5	5.03	4807
5	6	8.02	4535
6	7	12.12	4334
7	8	14.68	4051
8	9	23.04	3533
9	10	49.31	3158

- Table 2 -

6. REFERENCES

[1] Broeckx,F., L. Kaufman and M. Steleman, THE OPTIMAL LOCATION OF A SYSTEM OF PUBLIC FACILITIES, Proceedings of the International Conference on Systems Modelling in Developing Countries, Bangkok, 1978

[2] Benders,J.F., PARTITIONING PROCEDURES FOR SOLVING MIXED-VARIABLES PROGRAMMINC PROBLEMS, Numerische Mathematik, Vol 4, PP.238-252, 1972

[3] Revelle,C, D. Marks, J.C. Liebman, AN ANALYSIS OF PRIVATE AND PUBLIC SECTOR LOCATION MODELS, Man. Science, Voi 16, NO 11, PP. 692-707, 1970

[4] Hakimi,S.L, OPTIMUM LOCATIONS OF SWITCHING CENTERS ANC THE ABSOLUTE CENTERS AND MEDIANS OF A GRAPH, Opns. Research, Vol 12, PP.450-459, 1964

[5] Garfinkel and Neebe, AN ALGORITHM FOR THE M-MEDIAN PLANT LOCATION PROBLEM, Transp. Sc., Vol 8, NO 3, PP. 217-236, 1974

[6] Baker, A HEURISTIC APPROACH TO LOCATING A FIXED NUMBER OF FACILITIES, Log. and Transp. Rev. (Canada), Vol 10, NO 3, PP.195-201, 1974

[7] Järvinen,P, J. Rajala and H. Sinervo, A BRANCH-AND-BOUND ALGORITHM FOR SEEKING THE p-MEDIAN, Opns. Research, Vol 20, PP. 173-178, 1972

[8] Balinski, M.L., ON FINDING INTEGER SOLUTIONS TO LINEAR PROGRAMS, IBM Data Processing Division, PP. 195-292, 1964

[9] Sa,G, BRANCH AND BOUND AND APPROXIMATE SOLUTIONS TO THE CAPACITATED PLANT LOCATION PROBLEM, Opns. Research, Vol 17, PP. 1005-1016, 1969

[10] Davis,F.S., T.L. Ray, A BRANCH-BOUND ALGORITHM FOR THE CAPACITATED FACILITIES LOCATION PROBLEM, Naval Res. Log. Quat., Vol 16, PP. 331-343, 1969

[11] Geoffrion,A and G. Graves, MULTICOMMODITY DISTRIBUTION SYSTEM DESIGN BY BENDERS DECOMPOSITION, Working paper NO 209,Western Management Science Inst., UCLA, 1973

[12] Van Oudheusden,D and F. Plastria, UN MODELE DE LOCALISATION COMME APPLICATION DE LA THEORIE DES JEUX, Internal Report, CSOOTW112, University of Brussels, 1978

OPERATIONS RESEARCH APPLICATIONS TO TOURISM*

W. W. SWART TURGUT VAR CHARLES E. GEARING

ABSTRACT. This paper categorizes the application of Operations Research to Tourism as falling into the areas of tourist forecasting, the determination of tourist flows, the measurement and evaluation of the impact of tourism oriented facilities and activities, and the modelling of decision planning and policy problems. Each area is then further subdivided - i.e. tourism forecasting into time series analysis, causal, and qualitative approaches; the evaluation of impact into classical economic approaches, derived value approaches, surrogate approaches, and a longitudinal approach; and decision models into prescriptive and descriptive approaches. Within each subdivision, the pertinent findings obtained by a review of the literature for Operations Research oriented studies applied to that subdivision are reported. It is concluded that a greater awareness of each others area by researchers in both fields will have the potential of providing synergistic results in future work.

---

*Forthcoming, Annals of Tourism Research, Winter 1979.

INTRODUCTION

Operations Research (OR) is a term which stands for an approach to problem solving characterized by a systems orientation, an interdisciplinary philosophy, a focus on the quantification of the relevant aspects of the situation into a model, and the manipulation of this model through the use of mathematical, statistical and computer methodologies in order to develop decisions, plans, and policies.

The term was coined during the early days of WWII when teams of scientists of different disciplines were asked to pool their various areas of expertise into efforts to assist the military in arriving at solutions to various operational problems. These problems ranged from determining the best location of a limited number of radar units to give early warning of air raids to determining search strategies for submarine detection.

Since then, OR has become a standard activity within most organizations and is taught as a part of all business and many engineering curricula under such varied names as management science, quantitative analysis, systems analysis, and others. The literature in these areas exhibits many applications of Operations Research attesting to its increasing importance which can be attributed to the widespread availability and use of computers in today's world.

Within the field of tourism, the term OR appears to be first used in the literature by Cesario (11). However, then as today, many tourism researchers and planners use concepts and methodologies which, in the business world, are often associated with the practice of Operations Research. These concepts and methodologies most often deal with problems in the areas of: tourist forecasting; the determination of tourist flows; the measurement and evaluation of the impact of tourism oriented facilities and activities; and the modelling of decision, planning, and policy problems.

In the subsequent sections of this paper the use of Operations Research in each of the problem areas listed above will be surveyed. In all cases, the determination of whether the papers included in the survey can be considered as Operations Research has been made subjectively based on the consensus and experience of the authors of this article.

TOURIST FORECASTING MODELS

In this section, approaches to determining the number of tourist arrivals to a touristic region are considered. The problem of predicting the flows of tourist between one touristic attraction and another within the same geographical region is not included in this section since the methodology appropriate for this analysis can be different and will be presented in the next section.

From an Operations Research viewpoint, forecasting models are often classified as being causal, time series, or qualitative in nature. Causal models are those which explicitly attempt to quantify the relationship between a set of causal, or independent variables, and the variable of interest. All regression and econometric models fall into this category. Time series models are not concerned with explaining the reason why the forecast is what it is (as causal models attempt). Instead, all causal factors are considered in the aggregate by the assumption that the net result of these variables is what has caused whatever trends, seasonalities, and cyclical behavior may exist in the data, and that an extrapolation of the trend, seasonal, and cyclical behavior will yield an appropriate forecast.

In contrast with the above two approaches which require the existence of past data, qualitative forecasting models are designed to elicit and capture the consensus of experts in the field of interest with regard to what the future will bring. This approach is particularly useful when forecasts are to be made for a distant point in the future, or for situations which are unique and novel.

In the next sections, applications of each of the model categories will be briefly discussed.

Time Series Analysis

An excellent discussion of the nature, impact, and importance of seasonality in tourism has been provided by BarOn (4,5). In these two publications, BarOn describes the use of the time series analysis program "X-11", developed by Shiskin at the U.S. Bureau of the Census, to quantify the trend, seasonal, cyclical components of several time series associated with the tourism industry in Israel. Specifically, the following time series are analyzed in BarOn (4):

a) Tourists arriving and departing by air

b) Tourists arriving

c) Foreign currency-income from tourism in Israel

d) Residents departing

e) Bed nights in tourist recommended hotels broken down into visitors from abroad, Israelis, and totals.

In BarOn, (5), these series were compared to those of several other countries. Furthermore, several examples are given regarding how the results can be used as input to and guidelines for planning studies.

In another application of time series forecasting methods, Geurts and Ibrahim (24) apply two well-known approaches, namely the Box-Jenkins technique and Brown's exponential smoothing model, to tourist arrival data for Hawaii. The main purpose of this article is to compare the two approaches, and the results indicate that for the Hawaii tourist arrival data, the Brown exponential smoothing approach is preferable to the Box-Jenkins technique, because of its lower cost, although the accuracy is not particularly superior.

The Canadian Government Office of Tourism, in a report prepared by Liepa and Chau, (27) has also published the results obtained when the Box-Jenkins technique was applied to eight time series of importance to the tourism industry. Four of these series were travel accounts and consisted of quarterly data running from the first quarter of 1959 to the first quarter of 1976. The conclusions to the report are rather tentative, but indicate that the Box-Jenkins technique may have great potential for a variety of

forecasting applications.

Casual Models

The Battelle Research Centre in Geneva, Switzerland (2) has developed one of the more complete causal forecasting models for international tourist flow reported in the literature. The model is stepped since it begins by hypothesizing a simple relationship between the number of tourists annually generated to one country by another country and the latter's population; then a number of other independent variables such as per capita income, language similarity, attractiveness, etc., can be progressively introduced in order to obtain the most satisfactory outcome.

Crampon and Tan (18), propose a model which is conceptionally similar to the Battelle model, except that the hypothesized causal factors are more specific to tourism in the Pacific.

Cline (17), has proposed a more traditional multiple regression model in which the causal variables are primarily economic in nature. The results of this model, which was developed in conjunction with the Midwest Research Institute, provide forecasts of future travel growth.

The extensive efforts by Parks Canada, which were started in 1967, in forecasting the demand for outdoor recreation (the CORD studies) have been thoughtfully documented by Beaman, Heit, and Do (8). The entire thought process involved with this study, which resulted in an extensive series of technical notes, is reviewed, and the shortcomings of many classical approaches and models are brought out in this presentation. Furthermore, the importance of behavioral, as well as socio-economic, considerations are continually emphasized. Although primarily a demand estimation study, many of the methodologies developed utilize measures of attractiveness including travel distance as causal variables. The attractiveness considerations, which are also considered in a recreational demand study by Tapiero (41), and another study by BarOn and Schechter, (6), will be addressed in a subsequent part of this paper.

Qualitative Models

Although both the Battelle Model and the Crampon model contain subjective factors, as do most other traditional

flow models surveyed in Gearing, Swart, Var (21), their basic nature is that of a causal model in which the subjective factor is expressed as an index. (Some recent approaches to obtaining this index are discussed later).

A procedure suggested by Gaumnitz, Swinth, and Tollefson (19), which is primarily qualitative, illustrates how consensus seeking and group decision making methods can be of use. The approach is based on methodology similar to that presented in the OR literature and involves the concept that an action such as the visit to some lake by a person or family is assumed to be the result of a decision. Such a decision is a mental cognitive process. That is, it is a conscious mental consideration and evaluation of the characteristics of various possible courses of action in the light of one's needs, goals, and limitations (e.g. funds, time, skills, etc.). Accordingly, it should be possible to characterize the decision processes of a person in the form of a program. One ought to be able to specify a series of statements, questions, or commands which, if followed, will produce choices that match those of the individual. This program, or discrimination net, can then be helpful to a planner/manager in predicting actual choices of potential visitors, and hence, usage rates or forecasts. Of course, this forecast is contingent on an analysis of the structure of the discriminant nets of a sample of individuals to the general client population.

Another qualitative model that has been used in various studies is the Delphi technique. It is a relatively complex technique originally developed to structure discussions and summarize opinions of a group without meetings and direct discussion. Its development and use as a technology assessment tool has been well documented. (Linstone and Turoff, (28)).

Moeller, Shafer, and Getty, (32), in order to predict leisure environments of tomorrow utilized the Delphi technique. They started with 904 experts and ended, after successive rounds, with 405 experts. Their study proceeded through four rounds. In round one, experts were asked to list the most significant events they felt would have a 50-50 chance of occurring by the year 2000, considering only events that related to their own area of expertise. Panel members were instructed to assume an essentially stable political

situation and sustained economic growth to the year 2000. By using four rounds they were able to predict the probabilities of future events associated with natural resource management, wildland-recreation management, environment pollution, population-workforce-leisure, and urban environment. Though some of the predictions projected to the year 2050 may sound fantastic now, the authors think that some of the events predicted may occur even sooner than forecast.

The second and relatively recent study that employed the Delphi technique was conducted for the Canadian Government Office of Tourism by L. J. D'Amore & Associates Ltd. in 1976. This study named Tourism in Canada-1986 (20) used two rounds. In round I of the Delphi survey, scenarios were introduced based on preliminary research conducted by the Bureau of Management Consulting, Informetrica, Bell Canada and L. J. D'Amore & Associates Ltd. These were intended to elicit qualitative and quantitative feedback from Round I panelists who were distributed into four separate panels: general, demand, resource, and impact.

In the second round the members of the general panel were distributed among the demand, resource and impact panelists. For the results of this round, respondents were asked to rate the probability of a scenario becoming reality by 1986. They were then invited to modify their scenarios to give them higher probabilities, or to provide other scenarios if they did not agree with any of those stated. In all cases, the scenarios in these questions were constructed from the composite responses of panelists in Round I.

TOURISM FLOW MODELS

The methodology for predicting tourist flows between the various touristic attactions in a geographical area can be different from that of predicting tourist arrivals in that knowledge of the number of tourists at one given attraction can be used in some cases as information in determining how many will be visiting another location within that area at a subsequent time. Although the state of the art in the application of flow models to tourism is not well developed, it appears that the major approaches can be divided into stochastic (probabilistic) and deterministic.

Probabilistic Approaches

A specific example to illustrate the stochastic approach is given by Mednick (31), who has developed a model designed to describe the probabilities that U.S. visitors to the province of Ontario would make overnight stops among the ten economic regions of the province. The model is a Markov chain whose transition probability matrix illustrates the relationships of the economic regions as locations of overnight stopping points. The results of the model yield the average number of overnight stops and their distribution through the province for visitors first stopping in each region of Ontario. Although apparently the first published application of Markov chains to predicting travel flows, this model has shown good results and sets the stage for further applications of this promising approach.

Deterministic Approaches

Cesario (14), presents a family of models and associated computational algorithms to estimate outdoor recreation flows in southern Ontario and discusses some practical uses of the models for policy making. The general approach taken falls into the category of causal models, and the causal factors are some of the attractiveness concepts developed as part of the CORD studies (12).

Gearing, Swart, and Var (22) developed a tourism flow model as part of a multiperiod planning model for tourism development. Their model differs from others in that it is primarily concerned with predicting changes in tourist flows which occur as a result of the implementation of specific tourism development projects in various locations. The basic hypothesis is that the development of new projects changes the "touristic" attractiveness of a region and that, as a consequence, existing tourist flows will be redistributed to reflect this change. The concept of "carrying capacity" is explicitly incorporated into the model by a scheme which directly relates the potential changes in touristic attractiveness resulting from the development of new projects to the "saturation index" of the area (saturation index is the ratio of projected visitors to the visitor capacity). The closer the saturation index is to it, the less effect any further development

can have on the attractiveness of the area, and hence on the tourist flow to and from that area.

EVALUATING THE IMPACT OF TOURISM

One of the primary motivators behind the consideration of developing tourism potential is the expectation that, once developed, an income stream will be generated which can then be used to enhance the region's other economic sectors. As such, the measurement of this income stream is perhaps the most direct, and certainly the most traditional, measure of tourism impact. In more recent work, the need to consider the social, cultural, and political impact of tourism has been stated. However, the state of the art is such that there are only a few documented cases where the quantitative evaluation of these impacts has been successful.

The approaches to evaluating the impact of tourism appropriate to consider in this paper can be loosely categorized as 1) those based on classical economic theories (e.g. input-output analysis, econometric, etc.) 2) those which do not deal with revenue streams (i.e. primarily those dealing with state parks, etc.), but instead imply these by using a measure of the "willingness to pay" or the value derived from the usage of these facilities, 3) those which use other surrogates for implying economic worth; and 4) those which use a longitudinal, or systems, approach to evaluating impacts.

"Classical" Economic Approaches

Powers (36) describes the ST method, a procedure now in use at the World Bank and the Inter-American Bank on an epxerimental basis. This method proposes three stages in project analysis: 1) a financial appraisal using market prices for all inputs and outputs inclusive of taxes and debt transactions arising from the project; 2) an economic efficiency analysis which prices inputs and outputs at their scarcity value to the economy as a whole; and 3) a social analysis which accounts for the distributional consequences of the investment.

Bargur and Arbel (3) have developed an objective function for their mixed integer programming model for comprehensive planning of Isreal's tourism industry (to be

further discussed in a subsequent Section 3) which reflects the contribution to total net income in foreign currency resulting from tourism activities.

Jud and Krause (25) have developed an econometric model to estimate the contribution of the visitor industry to economic growth in Puerto Rico. Additionally, they examined the externalities of tourism growth and attempted to delineate the major social costs imposed by an expanding tourist industry.

Powers (35) points out that, in addition to the benefits described above, the domestic resources savings due to the reduced marginal cost incurred, as well as the additional tourist services made available to domestic visitors, must be considered as benefits from international tourism projects.

In addition to the direct benefits of development in the tourism sector, the indirect benefits should also be taken into account. Archer (1) discusses how to calculate the multiplier effects of tourism spending. He also discusses the pattern of the initial round of tourist expenditures, examines the composition of the indirect and induced flows to see which sectors of the economy benefit from the multiplier effect, and presents the contributions multiplier analysis can make toward policy making and planning. His findings are supported by Shulz (41) who, in addition, criticizes many of the past methodologies used to calculate multipliers. He found that an estimate of five to ten times the true value of the multiplier is often used, leading to unfulfilled expectations about the potential benefits of tourism.

"Derived Value" Approaches

When evaluating benefits associated with investments which do not result in direct economic benefits (e.g. national parks, state recreation areas, etc.), the measurement methodology becomes much more subjective and a source of controversy. Knetsch and Davis (26), and Cesario (11), support methodologies which take the point of view that the primary benefits of recreation are those accrued by the user who is willing to give up a certain amount of his income in the pursuit of outdoor recreation.

The fact that the user is willing to give up something means that the recreational experience is of value to him. It is this implicit willingness to pay that provides clues as to the worth of the recreational site.

On the other hand, Mack and Myers (29) contend that it is not possible to assign a price to recreational benefits that is comprehensive and reliable. They suggest that investments by governments in recreation should be made so as to maximize the resulting social welfare. As a measure of social welfare they suggest the merit weighted user days of recreation that can be gotten in return for investing in recreation. The merit weights are assigned on the basis of the policy, conservation, and priority considerations supported by the government.

The current research thrusts in this area are provided by Parks Canada within the CORD Studies. In 1967 a series of joint Federal-Provincial research projects, known as the Canadian Outdoor Recreation Demand Studies, were initiated to provide information on recreation participation among Canadians and to discern the patterns and trends in the usage of outdoor recreational facilities by Canadians and by visitors.

While data collection efforts and rudimentary tabulation has been a routine function of park and recreation agencies, the factual basis for planning and policy decisions has remained poor. The aim of the CORD studies is to improve data collection and to make greater use of such information in guiding programs and investments. Examples of the CORD effort can be found in Cheung (15), Ross, (37), Cesario, (12) and Beaman, (7).

Another study carried out in Canada by the Bureau of Management Consulting (an official branch of the Government) deals with the development of a tourism impact model. In summary, the approach recommended by this study includes a systems model (with provincial components), integrated development of "delphi panels" for future forecasting, and social indicators to fill out the performance indicator aspects of a systems model. This model is intended to provide pragmatic assistance to policy makers who are faced with guiding the evolution of tourism (42).

Surrogate Approaches

Another approach toward measuring the benefit associated with the development of tourist accommodations and attractions is the development of a utility measure which reflects the touristic attractiveness of a given area. Gearing, Swart, and Var (23) developed an approach for the quantification of the notion of touristic attractiveness by constructing a multi-attribute utility function. This function incorporates weights derived by a procedure similar to that suggested by Churchman and Ackoff (16). The approach required that the following be determined:

(1) The criteria by which touristic attractiveness is judged, and

(2) the relative importance of this criteria due to another, as indicated by a series of numerical weights.

Then, with these two requirements satisfied, it was possible to

(3) employ the judgements of experts in making evaluations against these critieria and, using these inputs, to

(4) compute a numerical measure of the "relative attractiveness" of the touristic area. This, then, is the utility value associated with the site.

The results of the procedure were used to support the planning activities of the Turkish Ministry of Tourism. In particular, a ranking was obtained for the 65 geographical areas in Turkey which have been defined as touristic areas. The numerical utilities constituted, as it were, an "inventory" or assessement of the state of things in Turkey viz-a-viz the tourist. In addition, these concepts were modified so that it was possible to estimate a net increase of touristic attractiveness that would result as a consequence of carrying out a specific development project. This change in touristic attractiveness was then used as a surrogate measure for net foreign exchange earnings that would result from that project's development. Specifically, the development of an investment plan that would maximize

the increase in touristic attractiveness was taken to be the same as an investment plan that would maximize net foreign exchange earnings in the long run.

Nijkamp (33) approaches the problem of estimating the impact of tourism in three stages. In the first stage, a demand/supply analysis of the tourism sector is developed by specifying a set of criteria which constitutes a tourist profile. Each touristic region is then rated on a 0-10 scale as to the extent each criterion is demanded and available. The tourist demands and regional supplys are then superimposed and the most probable choice of tourist areas by tourists is obtained via generalized distance analysis.

In the next stage, a model is developed which considers the influence of regional characteristics, including the characteristics of the demand site itself and the tourist carrying capacity of that site, on tourist flows. This information is then made part of a modified input output analysis which, in addition to the demand for tourism commodities, indicates the way in which environmental attractiveness and tourist accommodations determine the total final demand of the tourism sector.

Bond and McDonald (10) are in the process of developing a surrogate impact evaluation scheme, referred to as "Tourism Barometers" which is designed to monitor relative changes in the tourism activity level of the State of Arizona in a quick and expedient manner so as to provide industry and government enough response time to react to changes in tourism trends and patterns.

A Longitudinal Approach

The approaches discussed so far in this section are all one time studies conducted at the planning stage of a project. Beaman, Lehtiniemi, and Stanley (9) point out that one common problem in assessing impact of tourism facilities is that those making the impact study, as well as the users of the study, often do not take into account the dynamic nature of the planning process. In other words, a tourism development is usually assessed in the early planning stage, and plans change as they advance, often as a consequence of the predicitons made. Hence, the final facility built may have little semblance to the facility

on which the initial impact prediction was based.

To remedy the above problems they provide a general planning framework which stresses the need for the establishment of a continuing interplay between impact analysts and park planners and developers. Their planning framework, in addition to the traditional concerns regarding the selection of data bases, projection procedures, gathering of data, etc. includes concerns about specification of objectives and development options and the identification of variables the monitoring of which will indicate whether the facility is developing in a way consistent with objectives.

DECISION MODELS

In the prior two sections, some of the approaches that have been used to predict the future in terms of number of tourists and their flows have been reviewed. The information generated by the forecasting activities is often used when the benefits associated with a specific tourism project are to be estimated by the approaches reviewed in Section IV, or other approaches. However, many situations encountered by the tourism planner involve not only the evaluation of one specific alternative, but the broader problem of selecting the best set of alternatives from among a number of possible choices each having different characteristics and resource requirements, as well as benefits.

For problems such as the above, the structuring of the characteristics of the decision problem together with the specification of the appropriate evaluation methodology for the possible alternatives into a decision model - i.e. a mathematical, statistical or logical representation - is often useful for then the power of the computer can be used to reduce the often burdensome computations regarding tradeoffs among alternatives, etc.

Developing decision models is the essence of Operations Research and is one of the characteristics that distinguishes OR from other quantiatively oriented disciplines. Decision models can broadly be categorized into two overlapping classes: prescriptive and descriptive. Prescriptive models are those whoch yield an answer which prescribes a "best" course of action for the decision problem modelled. Descriptive models, on the other hand, are those which

describe, or evaluate, the effect of a decision which was postulated by the user of the model. In the next section examples of each model type which have been applied to tourism problems will be reviewed.

Prescriptive Models

One of the first decision models used in the tourism field was designed to provide the input for the tourism sector into a five year national economic plan for Turkey. As such, the model was primarily designed to develop a tourism investment planning policy as a function to the budget allocated to that sector.

The basic model, reported in Gearing, Swart & Var (21), considered the 65 touristic planning areas of Turkey as each containing a well defined and interrelated set of investment alternatives. Each alternative required a given investment, and, if approved, would enhance the region's "touristic attractiveness" by a given amount. The model then selected that set of projects which would maximize the increase of Turkey's touristic attractiveness without exceeding the budget allocated to investments into the tourism sector and without exceeding other technological and practical limits. A "second generation" version of this model has been developed to include considerations such as multiple planning periods and carrying capacities of touristic areas and is reported in (22).

Another example of the use of prescriptive models is the comprehensive planning model of the tourist industry for Israel developed by Bargur and Arbel (3). It may be described as follows:

The objective of the model is to maximize the total net income in foreign currency from tourist activities.

This objective is subject to the following considerations and restrictions:

(a) The economic efficiency criterion - the cost of a unit of net value added of foreign exchange.

(b) Supply constraints of local production factors (accommodation facilities, manpower, and recreational sites).

(c) Demand constraints for the planning horizon based on demand forecasts which were developed exogenously under

varying assumptions.

The analysis resulting from the application of the model is designed to provide satisfactory answers to the following facets of planning:

(a) Determination of feasible and optimal level of output, and its mix from the viewpoint of the national economy.

(b) Balancing optimal supply and demand and the identification of surplus or deficit within the above categories.

(c) Determination of required inputs necessary to satisfy the optimal levels of output in terms of major production factors.

(d) Determination of the regional distribution of the various activities subject to the touristic potential of each specific region.

(e) Determination of the feasible and preferred seasonal distribution of the various activities.

(f) Realization of the shadow prices, i.e. the marginal values of changes in basic assumptions, input data, and optimal results.

Within the context of tourism planning, McMahon (30), presents a prescriptive model for the selection of tourism development projects in Puerto Rico, subject to limitations in funds and the quantity of skilled manpower available.

In a different context and of different scope, Saitta and Schnederman (38) propose a model to aid managers of individual state parks in planning the expansion of their facilities so as to satisfy the ever increasing user demands. The solution of the model provides information which indicates the number of additional recreational facilities which could be built under varying amounts of capital investment, maintenance costs, land requirements, and overall state recreation requirements.

Penz (34), formulates a linear programming model designed to indicate appropriate long-range visitor admittance policy. In addition, he indicates how, through selection of the appropriate objective function, the same model might serve to provide insights for internal park control. The model itself incorporates constraints

indicating the capacity of the park for visitors as well as constraints to model the transition of visitors between facilities within the park.

Descriptive Models

All of the models discussed in the prior sections fall into the category of prescriptive or optimization models in the sense that the results provided by the model indicate a "most desirable" course of action to be pursued. The principal drawback with the use of these prescriptive models is that they are not flexible enough to lend themselves to a number of possible situations. When this is the case, the use of descriptive models to aid the decision maker in evaluating alternatives which he formulates can be of great value. The most general type of descriptive model is a simulation model, and Cesario (3), provides several potential uses of this technique in the planning of outdoor recreation facilities.

An actual application of the use of simulation in tourism planning is presented by Say (39). The objective of his study is to develop a general planning model for the tourism sector of Turkey that will predict its path of development under alternate decision courses provided by the Ministry of Tourism. In particular, the model simulates the development of the tourism sector in future years and estimates the required amounts of capital, the necessary increases in debt capacity, and the resulting economic benefits that will be accrued.

CONCLUSIONS

This paper has provided an identification of the areas of concern to tourism professionals in which Operations Research has been applied. In addition to providing a brief review of some specific applications, the purpose of the paper is to suggest that Operations Research is a field of knowledge which can contribute methodologically to the study of tourism problems, while tourism is an area which can offer a fertile field of application for Operations Research methodologies and, consequently, that a great potential for a synergistical development of both areas exists.

BIBLIOGRAPHY

[1] Archer, H., "The Uses and Abuses of Multipliers", in Gearing, C., Swart, W., and Var, T. PLANNING FOR TOURISM DEVELOPMENT: QUANTITATIVE APPROACHES, Praeger Publishers, New York, N.Y., 1976

[2] Armstrong, C. W. G., INTERNATIONAL TOURISM: COMING OR GOING, Futures, June, 1972

[3] Bargur, J., and Arbel, A., A COMPREHENSIVE APPROACH TO THE PLANNING OF THE TOURISM INDUSTRY, Journal of Travel Research, Vol. XIV, No. 2, Fall, 1975

[4] BarOn, R. R., SEASONALITY IN TOURISM - PART 1, International Tourism Quarterly, Special Article No. 6, 1972

[5] ___________ , SEASONALITY IN TOURISM - PART 2, International Tourism Quarterly, Special Article No. 6, 1973

[6] BarOn, M. and Schechter, N., SIMULTANEOUS DETERMINATION OF VISITS TO A SYSTEM OF OUTDOOR RECREATION PARKS WITH CAPACITY LIMITATIONS, Regional and Urban Economics, Volume 3, No. 4, 1973

[7] Beaman, J., DISTANCE AND THE 'REACTION' TO DISTANCE AS A FUNCTION OF DISTANCE, Journal of Leisure Research, Volume 6, SUM, 1974

[8] Beaman, J., Heit, M. and Do, H., ESTIMATING PARTICIPATION IN OUTDOOR RECREATION ACTIVITIES: INSIGHTS GAINED FROM THE CORD STUDY RESEARCH. Socio-Economic Research Division, Program Coordination Branch, Parks Canada, Ottawa, Ontario, Canada, 1977

[9] Beaman, J., Lehtineimi, L. and Stanley, R., EVALUATION OF THE IMPACT OF A NATIONAL PARK: THE LONGITUDINAL STUDY. Socio-Economic Research Division, Program Coordination Branch, Parks Canada, Ottawa, Ontario, Canada, 1978

[10] Bond, M. E., and McDonald, W., TOURISM BAROMETERS: THE ARIZONA CASE. Paper presented at the Joint Meeting of the Operations Research Society of America and the Institute of Management Sciences, Atlanta, Georgia, November 6-9, 1977

[11] Cesario, F. J., OPERATIONS RESEARCH IN OUTDOOR RECREATION, Journal of Leisure Research, Volume 1, (Winter, 1969), No. 1

[12] ______________, FINAL REPORT ON ESTMATING PARK ATTRACTIVENESS, POPULATION CENTER EMMISSIVENESS AND EFFECTS OF DISTANCE (LOCATION) TN OUTDOOR RECREATION TRAVEL. Ottawa: Department of Indian and Northern Affairs, 1973, CORD Technical Note No. 4 (mimeo)

[13] ______________, A SIMULATION APPROACH TO OUTDOOR RECREATION PLANNING, Journal of Leisure Researc Vol. 7:1

[14] ______________, COMPARISON OF MODELS FOR ESTIMATING TRAVEL, TOURISM AND RECREATION FLOWS. Cornell University, Mimeographed paper, 1976

[15] Cheung, H. K., A DAY-USE PARK VISITATION MODEL, Journal of Leisure Research, Volume 4, (Spring, 1972)

[16] Churchman, C. W. and Ackoff, R. L. and Arnoff, INTRODUCTION TO OPERATIONS RESEARCH. John Wiley & Sons, New York, N.Y., 1957

[17] Cline, R. S., "Measuring Travel Volumes and Itineraries and Forecasting Future Growth to Individual Pacific Destinations", in Ladany, S. (Edt.), MANAGEMENT SCIENCE APPLICATIONS TO LEISURE TIME OPERATIONS, American Elsevier Publishing Co., Inc., New York, N.Y., 1975

[18] Crampon, L. J. and Tan, K. T., A MODEL OF TOURISM FLOW INTO THE PACIFIC, Tourist Review, July-Sept., 1973

[19] Gaumnitz, J. E., Swinth, R. L. and Tollefson, J. O. SIMULATION OF WATER RECREATION USER'S DECISIONS. Paper presented at the Tims XX International Meeting, Tel-Aviv, Isreal, June 26, 1973

[20] D'Amore, L., THE SIGNIFICANCE OF TOURISM TO CANADA, The Business Quarterly, Autumn, 1976

[21] Gearing, C. , Swart, W., and Var, T., O/R and M/S IN TRAVEL AND TOURISM. Paper presented at the Joint ORSA and TIMS Meeting in San Juan, Purerto Rico, 1974

[22] Gearing, C., Swart, W., and Var, T., A MULTI-PERIOD PLANNING MODEL FOR TOURISM DEVELOPMENT. Paper presented at the TIMS XXII International Meeting, Kyoto, Japan, July 26, 1975

[23] ______________________________ , PLANNING FOR TOURISM DEVELOPMENT - QUANTITATIVE APPROACHES. Praeger Publishers, New York, N.Y., 1976

[24] Geurts, Michael D., and Ibrahim, I.B., COMPARING THE BOX-JENKINS APPROACH WITH THE EXPONENTIALLY SMOOTHED FORECASTING MODEL APPLICATION TO HAWAII TOURISTS, Journal of Marketing Research, Vol. XII, (May 1975)

[25] Jud, G. Donald and Krause, W., EVALUATING TOURISM IN DEVELOPING AREAS, Journal of Travel Research, Vol. XV, No. 2, Fall 1976

[26] Knetsch, J. L. and Davis, R. K., "Comparisons of Methods for Recreation Evaluation" in Kneese, V., and Smith, S. (Edts.), WATER RESEARCH. Johns Hopkins Press, Baltimore, Md., 1966

[27] Liepa, R., and Chau, P., METHODOLOGY FOR SHORT TERM FORECASTS OF TOURISM FLOWS. Research Report No. 4, Economic Research Section, Policy Planning and Industry Relations Branch, Canadian Office of Tourism, Ottawa, Canada, January, 1977

[28] Lindstone, H., and Turoff, M., (Eds.), THE DELPHI METHOD: TECHNIQUES AND APPLICATIONS. Addison Wesley Co., Reading, Mass., 1975

[29] Mack, R. and Myers, S., "Outdoor Recreation", in Dorfman (Ed.), MEASURING THE BENEFITS OF GOVERNMENT INVESTMENTS, Brookings Institution, Washington, D.C., 1965

[30] McMahon, J. J., TOURISM PLANNING AND PROJECT EVALUATION - THE PUERTO RICAN CASE. Paper presented at the ORSA/TIMS Joint National Meeting, San Juan Puerto Rico, October 16-18, 1974

[31] Mednick, H., A MARKOV CHAIN MODEL OF TRAVEL PATTERNS OF U.S. VISITORS TO ONTARIO. Journal of Leisure Research, Vol. 7:4, 1975

[32] Moeller, G., Shafer, E., and Getty, R., "Leisure Environments of Tomorrow", OUTDOOR RECREATION - ADVANCES IN APPLICATIONS OF ECONOMICS, U.S. Forest Service, U.S. Department of Agriculture, Washington, D. C., 1977

[33] Nijkamp, P., ENVIRONMENTAL ATTRACTION FACTORS AND REGIONAL TOURIST EFFECTS. Research Memorandum No. 14, Free University, Amsterdam, The Netherlands

[34] Penz, A. J., OUTDOOR RECREATION AREAS: CAPACITY AND THE FORMULATION OF USE POLICY, Management Science, Vol. 22:2, October 1975

[35] Powers, T. A., APPRAISING INTERNATIONAL TOURISM PROJECTS. Papers on Project Analysis No. 1, Inter-American Development Bank, Washington, D. C., July, 1974

[36] ____________, PUBLIC PLANNING FOR TOURISM IN DEVELOPING COUNTRIES. Paper presented at the XXIII International Meeting of TIMS, Athens, Greece, July, 1977

[37] Ross, J. H., A MEASURE OF SITE ATTRACTION. Ottawa, Environment Canada, Occasional Paper No. 2 (reprint), 1973

[38] Saitta, W. W. and Schedeman, I. W., NEW DIMENSIONS IN PARK MANAGEMENT THROUGH LINEAR PROGRAMMING. Journal of Leisure Research, Vol. 4, No. 4, 1972

[39] Say, A., A MODEL FOR TOURISM PLANNING IN TURKEY. Paper presented at the Joint National ORSA/TIMS Meeting, Atlanta, Georgia, November 7-9, 1977

[40] Shulz, R. A., THE TOURIST DOLLAR MULTIPLIER: MULTIPLYING THE MULTIPLIER? Paper presented at ORSA/TIMS meeting, 1975, in Las Vegas, Nevada

[41] Tapiero, . S., THE DEMAND AND UTILIZATION OF RECREATIONAL FACILITIES: A PROBABILITY MODEL. Regional and Urban Economics, Vol. 4, No. 2, October, 1974, pp. 173-186

[42] Var, T. and Knetsch, Jack L., MACRO ECONOMIC TOURISM RESEARCH IN CANADA. Paper presented at the International Association of Scientific Experts in Tourism

Conference, September, 1975, Belgrade, Yugoslavia

MULTICRITERIA DECOMPOSITION IN A
DECENTRALIZED ORGANIZATION

SANG M. LEE* and B.H. RHO**

*First National Lincoln Distinguished Professor of Management and Department Chairman, University of Nebraska, Lincoln, Nebraska 68588

**Senior Research Fellow, Korea Development Institute, Seoul, Korea

ABSTRACT. The decomposition approach, originally developed by Dantzig and Wolfe, has received increasing attention in recent years not only as a computational technique for large-scale management models but also as a systematic tool for designing organizational structure and information systems in decentralized organizations. One of the major deficiencies of the Dantzig-Wolfe approach is its inability to consider multicriteria decomposition problems due to its linear programming formulation. Perhaps the most effective technique that can handle multicriteria decision making problems is goal programming.

In this paper goal programming is applied to the decomposition process in order to facilitate multiple conflicting objectives in a decentralized organization. The development of the goal programming based decomposition model not only provides the capability of dealing with multiple objective problems but also a managerial implication in terms of evaluating organizational effectiveness of a decentralized organization. This implication is derived through the sensitivity analysis of a shadow price vector. A simple model formulation example is presented for an illustrative purpose.

MULTICRITERIA DECOMPOSITION IN A DECENTRALIZED ORQANIZATION

INTRODUCTION

Resource allocation in a decentralized organization has been an important research topic for economists, behavioralists, and management scientists. Excellent economic studies can be found in Arrow and Hurwicz [1], Malinvaud [19], and Burton, et al. [2]. Behavioral scientists have dealt with the decentralized decision process primarily on the basis of personal experience and empirical studies. Management scientists have addressed the problem in connection with mathematical decomposition techniques, as can be found in Dantzig and Wolfe [7], Charnes, et al. [3], Geoffrion [10], Grinold [11], Kornai and Liptak [16] Ruefli [21], and Jennergren [13]. Many of the above cited studies have been criticized for the lack of either analytical rigor as in the case of behavioralists or behavioral implications as in the case of economists and management scientists.

Decomposition analysis was originally developed by Dantzig and Wolfe [6] as a computational device for solving large scale linear programs. Recently, however, it has received increasing attention [2], [9], [14], [18], [21] because of two important characteristics: (1) the decomposition technique can be utilized for resource allocation in decentralized organization; and (2) it provides management in decentralized organizations behavioral insights for developing organizational structure and information systems. Many recent studies have attempted to integrate analytic and behavioral aspects of the decomposition process, as can be found in Kornai [15], Ruefli [21], Collomb [5], Freeland [8], and Freeland and Baker [9].

One of the major deficiencies of the Dantzig-Wolfe (DW) approach is its inability to consider multicriteria decomposition problems due to its linear programming formulation. Perhaps the most effective technique that can handle multicriteria decision making problems is goal programming [4], [12], [17]. In this paper goal programming is applied to the decomposition process in order to facilitate multiple conflicting objectives in a decentralized organization.

The development of the goal programming based decomposition model also provides a managerial implication in terms of evaluating organizational effectiveness of a decentralized organization. This implication is derived on the basis of the Kim's [14] methodology which defined the concept of "going concern value" through a sensitivity analysis of the shadow price vector.

THE MULTICRITERIA DECOMPOSITION MODEL

The model to be presented here is based on the basic Dantzig-Wolfe [6] cooperative scheme for a two level hierarchical organization. The organizational structure (see Figure 1) is based on the central unit as a superordinate and a number of divisions as subordinates. A two level organization is used in this study as it provides more simple conceptualization while describing the basic characteristics of the resource allocation decision process in a decentralized organization.

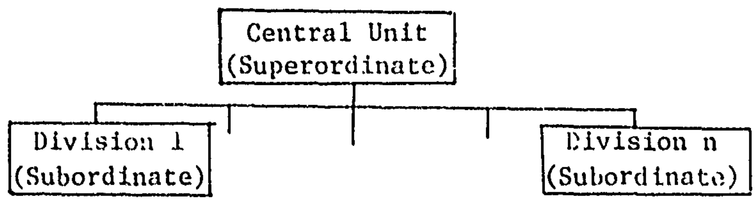


Fig. 1 A Two Level Organizational Structure

Two types of goals are involved in the model: (1) each division formulates goals which cannot be shared by other divisions, (2) the central unit formulates and allocates goals. The objective of the central unit is to assign corporate goals to divisions so as to attain them to the fullest possible extent.

The model is expressed as follows:

$$\text{Min } Z = \sum_{i=0}^{n} (P_i^- D_i^- + P_i^+ D_i^+) \tag{1}$$

$$\begin{array}{lllllll}
\text{subject to: } A_1X_1 + A_2X_2 + \dots + A_nX_n & + ID_o^- & - ID_o^+ & = b_o \\
B_1X_1 & + ID_1^- & - ID_1^+ & = b_1 \\
\quad B_2X_2 & + ID_2^- & - ID_2^+ & = b_2 \\
\quad\quad \ddots & \vdots & \vdots & \vdots \\
\quad\quad\quad B_nX_n & + ID_n^- & - ID_n^+ & = b_n \\
 & X_i, & D_i, D_i & \geq 0
\end{array}$$

where, X_i is $(n_i \times 1)$ vector of decision variables; A_i and B_i are the $(m_o \times n_i)$ and $(m_i \times n_i)$ matrices of technological coefficients of the central unit and divisions, respec-

tively; D_i^- and D_i^+ are (m_i x 1) vectors of deviational variables from goal i; P_i^- and P_i^+ are (1 x m_i) vectors of preemptive (ordinal) priority factors, with differential weights attached, associated with D_i^- and D_i^+, respectively; and b_o and b_i are (m_o x 1) and (m_i x 1) vectors of resources (or goals) of the central unit and divisions, respectively.

The algorithm decomposes only divisional units, and the central unit generates transfer prices of corporate resources. Each division, with these transfer prices in mind, then solves its own subproblem.

Let us define the following variables:

$$Y_i = \left[X_i,\ ID_i^-,\ ID_i^+\right] \text{ for } i=1, \ldots, n;$$

$$Q_i = \left[\bar{0},\ P_i^-,\ P_i^+\right] \text{ for } i=1, \ldots, n.$$

Then the model can be reformulated as follows:

$$\text{Min } Z = \sum_{i=1}^{n} Q_i Y_i + P_o^- D_o^- + P_o^+ D_o^+ \qquad (2)$$

$$\begin{aligned} \text{subject to: } & A_1Y_1 + A_2Y_2 + \ldots + A_nY_n + ID_o^- - ID_o^+ &= b_o \\ & B_1Y_1 &= b_1 \\ & \quad B_2Y_2 &= b_2 \\ & \qquad \ddots & \vdots \\ & \qquad\quad B_nY_n &= b_n \end{aligned}$$

Defining the extreme points of the convex polyhedron, $B_iY_i = b_i$, as Y_i^j ($j=1, \ldots, k_i$), we can express Y_i as a convex combination of these extreme points, namely,

$$Y_i = \sum_{j=1}^{k_i} \beta_i^j Y_i^j$$

where $\sum_{j=1}^{k_i} \beta_i^j = 1$ and β_i^j is a coefficient of convex combination for Y_i^j, the j^{th} extreme point of the convex set $B_iY_i = b_i$.

Thus, we can reformulate the original model into the master problem:

$$\text{Min } Z = \sum_{i=1}^{n} \sum_{j=1}^{k_i} Q_i Y_i^j \beta_i^j + P_o^- D_o^- + P_o^+ D_o^+ \qquad (3)$$

$$\text{subject to: } \sum_{j=1}^{k_1} A_1 Y_1^j \beta_1^j + \sum A_2 Y_2^j \beta_2^j + \ldots + A_n Y_n^j \beta_n^j + ID_o^- - ID_o^+ = b_o$$

$$\sum_{j=1}^{k_1} \beta_1^j = 1$$

$$\sum_{j=1}^{k_2} \beta_2^j = 1$$

$$\vdots$$

$$\sum_{j=1}^{k_n} \beta_n^j = 1$$

The decomposition algorithm can now be formulated as follows:
Let B = Basis Matrix

O_B = Vector of objective function coefficients of basic variables

$$H = \left[H_1, H_2, \ldots, H_i, \ldots, H_n\right]$$

$$\begin{bmatrix} A_1Y_1 & A_2Y_2 & \cdots & A_iY_i & & A_nY_n \\ 1 & 0 & & 0 & & 0 \\ 0 & 1 & & . & & 0 \\ . & 0 & & . & & . \\ . & . & & 1 & (m_o+1)^{th} \text{ row} & . \\ . & . & & 0 & & . \\ . & . & & . & & . \\ . & . & & . & & . \\ 0 & 0 & & 0 & & 1 \end{bmatrix}$$

$$T = \left[T_1, T_2, \ldots, T_{m_o} \vdots T_{m_o+1}, \ldots, T_{2m_o}\right]$$

$$= \begin{bmatrix} I & \vdots & -I \\ \bar{0} & \vdots & \bar{0} \end{bmatrix}$$

$$Po = \left[P_{o1},\ P_{o2},\ \ldots,\ P_{o},\ \ldots,\ P_{o,mo} \;\vdots\; P_{o,mo+1},\ \ldots,\ P_{o,2mo} \right]$$
$$= \left[P_o^- \;\vdots\; P_o^+ \right]$$

Step 1

Solve a set of n goal programming problems of the following form:

$$\text{Min } Z_i = Q_i Y_i - Q_B B^{-1} H_i$$

$$\text{subject to: } B_i Y_i = b_i \qquad i=1,\ldots,n$$

and compute

$$Z_{oi} = P_{oi} - Q_B B^{-1} T_i$$

Step 2

Determine $\zeta = \underset{i}{\text{Min}} \quad (Z_i,\ Z_{oi})$

If $\zeta \geq 0$, go to Step 4. Otherwise, go to Step 3.

Step 3

Introduce the i^{th} variable (β_i, D_o^- or D_o^+) corresponding to into the basic solution of the master problem. Then, obtain a new basic feasible solution. Go to Step 1.

Step 4

The optimum solution is identified as follows:

$$Y_i = \sum_{j=1}^{k_i} \beta_i^j Y_i^j \qquad \text{for } i=1,\ldots,n.$$

ILLUSTRATIVE APPLICATION PROBLEM

In order to illustrate an application of the multicriteria decomposition model, let us consider an organization with two divisions. Suppose the firm has two corporate resources: 100 units of skilled labor per month and a $5,000 monthly budget which will be distributed to two divisions. Division I produces two types of televisions, T_1 and T_2; and Division II, three types of radios, R_1, R_2, and R_3. Each division has its own resources which cannot be shared. Division I has 80 units of unskilled labor per month and 150 square

feet of factory space. Division II has 170 units of unskilled labor per month and 100 square feet of factory space.

It is also confirmed from a market study that T_1 and T_2 have minimum demand of 70 and 45 units per month, respectively.

Suppose each product has the following requirements and profit margins:

	T_1	T_2	R_1	R_2	R_3
skilled labor (units)	2	3	1	4	2
budget ($100)	1	2	1	2	3
unskilled labor (units)	2	1	5	8	12
factory space (sq-ft)	5	6	2	3	5
profit margin ($)	10	6	6	10	12

The company has set the following goals in the order of their priorities:

P_1: Avoid the overutilization of skilled labor, according to contract with the professional employees

P_2: Operate within the budget of $5,000

P_3: Minimize the underutilization of unskilled labor in accordance with the union contracts

P_4: Operate within the factory space limit

P_5: Meet the demand as much as possible (differential weights should be assigned according to the unit profit of each product).

Thus, the following goal programming problem is formulated:

$$\text{Min } Z = P_1 d_1^+ + P_2 d_2^+ + P_3(d_3^- + d_7^-) + P_4(d_4^+ + d_8^+)$$

$$P_5(5d_5^- + 3d_6^- + 3d_9^- + 5d_{10}^- + 6d_{11}^-) \qquad (4)$$

subject to:

$$2X_1 + 3X_2 + X_3 + 4X_4 + 2X_5 + d_1^- - d_1^+ = 100$$

$$
\begin{array}{l}
X_1 + 2X_2 + X_3 + 2X_4 + 3X_5 + d_2^- - d_2^+ = 50 \\
2X_1 + X_2 \qquad\qquad + d_3^- - d_3^+ = 80 \\
5X_1 + 6X_2 \qquad\qquad + d_4^- - d_4^+ = 150 \\
X_1 \qquad\qquad + d_5^- - d_5^+ = 70 \\
X_2 \qquad\qquad + d_6^- - d_6^+ = 45 \\
5X_3 + 8X_4 + 12X_5 + d_7^- - d_7^+ = 170 \\
2X_3 + 3X_4 + 5X_5 + d_8^- - d_8^+ = 100 \\
X_3 \qquad\qquad + d_9^- - d_9^+ = 80 \\
X_4 \qquad\qquad + d_{10}^- - d_{10}^+ = 70 \\
X_5 + d_{11}^- - d_{11}^+ = 90 \\
X_i,\ d_i^-,\ d_i^+ \geq 0 \\
P_k >>> P_{k+1}
\end{array}
$$

The optimum solution of the master problem of the above decomposition model obtained at the 10th iteration is as follows:

$$
X_B = \begin{bmatrix} \beta_1^9 \\ \beta_2^7 \\ \beta_1^2 \\ d_1^- \end{bmatrix} = \begin{bmatrix} 16/30 \\ 1 \\ 14/30 \\ 34 \end{bmatrix}
$$

$$Q_B = (20P_3 + 335P_5,\ 102P_5,\ 80P_3 + 485P_5,\ 0)$$

$$Q_B B^{-1} = (0,\ -2P_3 - 5P_5,\ 80P_3 + 485P_5,\ 68P_3 + 1198P_5)$$

$$Y_1^9 = \begin{bmatrix} 30 \\ 0 \\ 20 \\ 0 \\ 40 \\ 45 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad Y_1^2 = \begin{bmatrix} 0 \\ 0 \\ 80 \\ 150 \\ 70 \\ 45 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad Y_2^7 = \begin{bmatrix} 34 \\ 0 \\ 0 \\ 0 \\ 32 \\ 46 \\ 70 \\ 90 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Thus, the optimum solution to the original two divisional goal programming problem is:

$X_1 = 16 \quad X_2 = 0 \quad X_3 = 34 \quad X_4 = 0 \quad X_5 = 0$

$d_1^- = 34 \quad d_2^- = 0 \quad d_3^- = 48 \quad d_4^- = 70 \quad d_5^- = 54$

$d_6^- = 45 \quad d_7^- = 0 \quad d_8^- = 32 \quad d_9^- = 46 \quad d_{10}^- = 70$

$d_{11}^- = 90$

$d_i^+ = 0$, for all $i=1,2,\ldots,11$

$Z = 48P_3 + 1433P_5$

Thus, all goals are completely attained except the third and fifth goals.

MANAGERIAL IMPLICATION

Now, let us analyze the shadow priority vector $Q_B B^{-1} = (0, -2P_3 -5P_5, 80P_3 + 485P_5, 68P_3 + 1198P_5)$ of the optimum solution. The vector represents the so called "going concern value" of shared resources or divisional activities. The first two components represent the value of two corporate resources and the next two the value of two divisions. The interpretation should be evident from the following explanations.

If there is one unit increase in the budget, the underachievement of the goal will decrease by $(2P_3 + 5P_5)$. Let q_i be the change in the implementation weight of each proposal. Then, our marginal equation will be:

$$q_1 \begin{bmatrix} A_1Y_1^9 \\ \\ 1 \\ 0 \end{bmatrix} + q_2 \begin{bmatrix} A_1Y_1^2 \\ \\ 1 \\ 0 \end{bmatrix} + q_3 \begin{bmatrix} A_2Y_2^7 \\ \\ 0 \\ 1 \end{bmatrix} + q_4 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Substituting the values of Y_1^9, Y_1^2, and Y_2^7 into the previous equation:

$$q_1 \begin{bmatrix} 60 \\ 30 \\ 1 \\ 0 \end{bmatrix} + q_2 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + q_3 \begin{bmatrix} 34 \\ 34 \\ 0 \\ 1 \end{bmatrix} + q_4 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Thus, $q_1 = 1/30$, $q_2 = -1/30$, $q_3 = o$, $q_4 = -2$. These figures can be interpreted as follows: the implementation weight of the 9th proposal of Division I will be increased by 1/30 and that of the 2nd proposal of Division I will be decreased by 1/30. There will be no change in the proposal of Division II and d_1^- will be decreased by 2.

If we denote $C_i^j = Q_iY_i^j$, then $C_1^9 = 20P_5$, $C_1^2 = 80P_3 + 485P_5$, $C_2^7 = 1028P_5$. Then changes in the underachievement of the goal will be:

$$(1/30)\ (20P_3 + 335P_5) - (1/30)\ (80P_3 + 485P_5)$$

$$= -2P_3 - 5P_5$$

Now it should be evident that one unit increase in budget results in the improvement of goal achievement by $2P_3 + 5P_5$.

P_3 is the priority associated with underutilization of unskilled labor and P_5 that of demand, which can be interpreted as opportunity cost of unskilled labor and forgone demand, respectively.

"The going concern values" of Division I and Division II are $80P_3 + 485P_5$ and $68P_3 + 1198P_5$, respectively. Comparing these values with "idle values" [1] of $80P_3 + 485P_5$ for

1. The "idle value" is the value of a division when it is not engaged in production.

Division I and $170P_3 + 1130P_5$ for Division II, we may note that Division I contributes nothing to the organization as a whole by producing X_1 while Division II contributes $102P_3 - 68P_5$ by producing X_3. Division I's television production is capital intensive while Division II's radio production islabor intensive. Division I, in producing X_1, employs one unit of scarce budget resources and utilizes two units of unskilled labor while Division II, in producing X_3, employs one unit of budget and utilized five units of unskilled labor. For Division I, whatever contribution it would have made is cancelled out by paying for price of budget.

We may conclude that Division II is more effective in achieving organizational goals and achievement of organizational goals can be reduced by transferring unskilled labor units of Division I to Division II. For example, if 30 units of unskilled labor of Division I were transferred to Division II, the underachievement would be reduced from $48P_3 + 1433P_5$ to $30P_3 + 1445P_5$.

As noted above, the concepts of "going concern value" and "idle value" are important as the difference of these two is equivalent to the amount of contribution of a division toward achieving an organizational objective by engaging in the present activities. This "contribution value" is the basis for evaluating the effectiveness of a division toward achieving goals and for designing a new organizational structure to improve the organizational effectiveness. Thus, the multicriteria decomposition model provides important information for managing the decentralized organization by objectives as advocated by Drucker [7] .

SUMMARY

In this paper a multicriteria decomposition model is developed and an illustrative example presented. The model is capable of dealing with multicriteria decision problems in two level decentralized organizations. The model recognizes that in a decentralized organization conflicts may exist between levels and divisions concerning which objectives should be pursued. The model attempts to achieve organizational goals as closely as possible. That is, instead of trying to optimize one grandiose organizational objective criterion, a resource allocation program is sought to achieve multicriteria objectives as exemplified by Simon's [22] satisficing concept.

Another important property of the model is that it provides information that is important for managerial analysis. It is possible to identify "going concern value" and "idle value" of each division through the model. The "contribution value," which is the difference between "going concern value" and "idle value" is a good indication of the divisional effectiveness. Thus, it can serve as an important input to organizational design of a decentralized organization.

REFERENCES

[1] Arrow, K.J. and L. Hurwicz, "Decentralization and Computation in Resource Allocation," in Essays in Economics and Econometrics, R.P. Font, ed., University of North Carolina Press, 1960.

[2] Burton, R.M., W.W. Damon, and D.W. Loughridge, "The Economics of Decomposition: Resource Allocation versus Transfer Pricing," Decision Sciences, Vol. 5, No. 3 (July 1974).

[3] Charnes, A., R.W. Clower, and K.O. Kortanek, "Effective Control through Coherent Decentralization with Preemptive Goals," Econometrica, Vol. 35, No. 2 (1967).

[4] __________, and W.W. Cooper, Management Models and Industrial Applications of Linear Programming, Wiley, New York, 1961.

[5] Collomb, B.P., "Goal Interval Approaches to Interporal Analysis and Decentralization in Management," ph.D. Dissertation, University of Texas, Austin, 1971.

[6] Dantzig, G.B. and P. Wolfe, "Decomposition Principle for Linear Programs," Operations Research, Vol. 8, No. 1 (February 1960)

[7] Drucker, P., Managing for Results, Harper & Row, New Yrok, 1964.

[8] Freeland, J.R., "Conceptual Models of the Resource Allocation Decision Process in Hierarchical Decentralized Organizations," Ph.D. Dissertation, Georgia Institute of Technology, 1973.

[9] ______________, and N.R. Baker, "Goal Partitioning in a Hierarchical Organization," Omega, Vol. 3, No. 6 (1975).

[10] Geoffrion, A.M., "Primal Resource-Directive Approaches for Optimizing Nonlinear Decomposable Systems," Operational Research Quarterly, Vol. 18, No. 3(1970).

[11] Grinold, R.C., "Steepest Ascent for Large Scale Linear Programs, SIAM Review, Vol. 14, No. 3 (1972).

[12] Ijiri, Y., Management Goals and Accounting for Control, Rand-McNally, Chicago, 1965.

[13] Jennergren, L.P., "Studies in the Mathematical Theory of Decentralized Resource-Allocation," Ph.D. Dissertation, Stanford University, 1971.

[14] Kim, Chaiho, Introduction to Linear Programming, New York: Holt, Rinehart and Winston, Inc., 1971.

[15] Kornai, J., Mathematical Planning of Structural Decisions, Amsterdam: North-Holland, 1969.

[16] Kornai, J., and T. Liptak, "Two-Level Planning," Econometrica, Vol. 33, No. 1 (January 1965).

[17] Lee, S.M., Goal Programming for Decision Analysis, Auerbach, Philadelphia, 1972.

[18] ________, and B.H. Rho, "The Binary Search Decomposition Method," A paper presented at the Seventh Annual Meeting of Southeast America Institute for Decision Sciences, (February 1977).

[19] Malinvaud, E., "Decentralization Procedures for Planning," in Activity Analysis in the Theory of Growth and Planning, E. Malinvaud and M. Bacharach, eds., Macmillan, 1967.

[20] Rho, B.H., "Decomposition Goal Programming: Algorithms, Computations, and Managerial Implications," Ph.D. Dissertation, Virginia Polytechnic Institute and State University, 1976.

[21] Ruefli, T.W., "Analytic Models of Resource Allocation in Hierarchical Multi-level Systems," Socio-Economic Planning Science, Vol. 8 (1974).

[22] Simon, H.A., "On the Concept of Organizational Goal," Administrative Science Quarterly, Vol. 9, No. 1 (1964).

APPLICATION OF THE HIERARCHICAL PLANNING APPROACH TO REGIONAL DEVELOPMENT AND MANAGEMENT IN THE PHILIPPINES

JONA S. BARGUR and FELIX MERO

Tahal Consulting Engineers
54 Ibn Gvirol Street
Tel Aviv, Israel

ABSTRACT. The empirical application of a hierarchical planning approach to regional development and management, as applied for the Pampanga Delta and Candaba Swamp Project, the Philippines, is the subject matter of this paper. In this approach a three-level modeling system was developed to represent the planning levels of the Project, the sector and the regional macro-economy.

Each of the planning levels within the multi-level planning hierarchy employs a system of mathematical models to enable the quantification of a complex system of physical and socio-economic interdependencies. For the lower level, that of the project, a system of hydro- and agro-meteorological models was formulated and applied to derive water balances and flood peak frequencies under varying physical upstream and downstream conditions in the entire river basin. At the intermediate level, alternative development plans for the agricultural sector were subjected to a system of optimization models in order to obtain the preferred program of land uses, cropping patterns and water resource projects. Finally, at the highest level, an analytical regional macro-economic multi-sector Input-Output model was applied to check the program for consistency with respect to the regional macro-economy at the projected target year.

These models and systems are described briefly in this paper together with their interlinkages, and the main results of the planning system demonstrated in their hierarchical interrelationships.

1. THE PLANNING APPROACH

A comprehensive plan for regional development is usually characterized by a number of preference criteria, i.e. socio-economic, physical and environmental facets of the decision-making process should be accounted for within the overall development framework of the region.

The analysis of water resource systems on a regional basis has gained strong support in the last decade, in particular where the region encompasses one or more river basins.

In planning large-scale systems with a variety of technical, social, economic, and environmental considerations, the usual problems encountered in modeling and optimization are magnified and tend to confuse the analysis. This is due to the high dimensionality of the model in question, the large number of variables, and their complex interactions. Usually, most regional plans, as well as water resource development schemes, are subjected to numerous criteria for assessment, justification and implementation. In fact, almost every regional development program has its own set of competing or complementary goals and objectives. These multiple preference functionals seem to be a prerequisite for total comprehensive planning.

However, a programming framework with multiple objective functions is obviously rather difficult to deal with due to intractable mathematical complexities (Klahr [1]). In general, a simultaneous optimization of all separate functionals cannot be carried out by means of traditional programming techniques. Thus, alternative methods have been suggested and applied (Nijkamp [2], [3], [4]). The following procedure has been suggested by Nijkamp [4]:

(i) Assigning separately to each preference criterion a political weight which reflects the relative importance (trade-off) of each criterion. This procedure is limited due to the lack of reliable numerical values for the various political weights and this compels the planner to resort to a compromise selection of numerous solutions.

ii) Goal programming, i.e. identification of one objective function as a dominant decision criterion and imposition of adequate numerical constraints on the outcomes of the remaining functions. This enables the planner to reduce the problem to a uni-dimensional constrained programming model, but forces him again rather arbitrarily to assign numerical values for the limits of the other objectives.

(iii) Identification of a set of feasible efficient solutions. These are relevant if they give results which do not allow an increase in the value of one function without affecting the others. However, these efficient programming procedures do not guarantee a unique solution, though they do identify an efficiency frontier.

(iv) A somewhat compromising method - this will be the take-off point for this paper - is the hierarchical optimization method which is based on sequential optimization of the objective function according to a preassigned ordinal ranking of importance (Haimes [5], Mesarovic et al [6]).

Under this method, the set of constraints for each stage of the iterative process is co-determined by the optimal results obtained in the previous stages using the original constraints.

The hierarchical optimization approach seems to provide something of an answer to the limitations of the abovementioned methods, though it does present the planner with a formidable task, if he wishes to apply it to an actual real life problem, mainly due to the large dimensions involved and the lack of data.

It was recognized from the outset that a hierarchical optimization model should be applied to planning the development of the Pampanga Delta-Candaba Swamp Area of the Philippines for the 30-year planning period, i.e., for the year 2000, as defined in the Project Terms of Reference. However, it was also realized that it would be necessary, in order to cope with empirical complexities, to compromise and to relax to some extent the rigidity of this planning method. Thus normative or optimal solutions were bypassed at certain levels and replaced, within the hierarchical context of the planning approach, by positive models and simulated systems.

2. THE HIERARCHICAL MODELING SYSTEM

To satisfy the objectives of a multi-criteria regional plan, to resort to quantification without losing considerable resolution, and to overcome the complexities of empirical application, an iterative hierarchic modeling system, which would provide for an overall consistent marginal analysis of the regional planning spectrum, was adopted, formulated and operated.

In this hierarchical modeling system, a number of models are grouped together and interrelated in such a way that the output of one serves as the input to another more specific

model. The hierarchy is in the level of aggregation, i.e. the model at the lowest level generates technically feasible projects which are then screened at the intermediate level to ensure consistency at the highest level. This multi-level, or hierarchical approach thus provides for the decomposition of the large scale and complex regional system into "independent" sub-systems.

This approach enables the planner, or the system analyst, to analyze and comprehend the behavior of the sub-systems at lower levels and to transmit the information obtained to fewer sub-systems at the higher level. Each of the sub-systems is independently analyzed and different techniques of analysis applied, based on the nature of the sub-system, as well as on its planning objectives and constraints. The sub-systems are linked by decision variables manipulated at a higher level which enable an optimal, or at least consistent, solution of the whole system at the highest level to be obtained (Haimes et al [7]).

A conceptual illustration of a three-level hierarchical system, as applied to the project, is presented in Fig. 1.

The three-level hierarchical model is comprised of the Project Level, the Sectoral Level and the Regional Macro-Economy Level.

In this study, the project level was oriented towards the development of water resource projects which were simulated and analyzed by a system of hydro- and agro-meteorological models (see also Bargur [8]). The second level, that of the sector, dealt mainly with the agricultural sector. A system of optimization models for (1) allocation of land, (2) introduction of cropping patterns and (3) selection of the most efficient water resource projects from the projects identified at the lower level, was employed at this intermediate level.

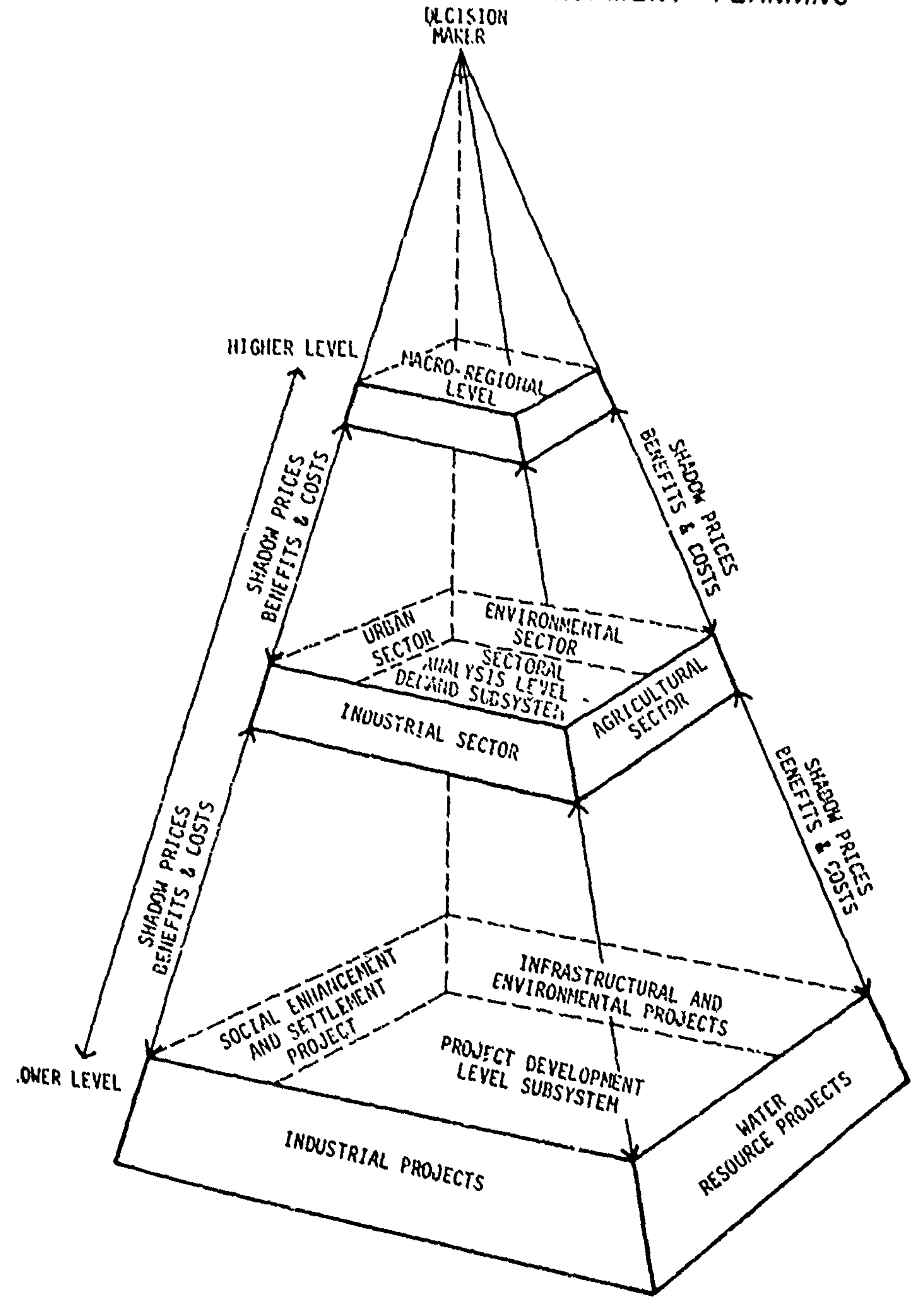
At the highest level of the hierarchy, the regional macro-economy was analyzed, employing a positive, projection model, which was aimed at testing the consistency of the results with respect to the agricultural sector, analyzed at the second level, within the overall macro-economy of the region. The structure and operational logic of the model enables a top down planning procedure as well as a bottom up approach.

The iterative nature of the analysis actually comprises a feedback planning process which ensures consistency and economic efficiency at all levels of the planning hierarchy.

After conclusion of the detailed analysis at each level, the planner can then synthesize the final and intermediate goals and link them to physical, environmental and economic constraints.

FIGURE 1

CONCEPTUAL ILLUSTRATION OF A MULTI-LEVEL(3) APPROACH TO REGIONAL DEVELOPMENT PLANNING



3. MULTI-LEVEL FUNCTIONAL INTERRELATIONSHIPS - FORMULATION OF THE HIERARCHICAL MODELING SYSTEM

3.1 Definition of the Inter-level Decision Variables

The essence of any hierarchical modeling system is the linkage between the various levels of the system in order to secure the necessary inter-level consistencies in the process of satisfying the overall objective of the development plan.

The planning levels or sub-systems are linked together by relevant decision variables. Each set of decision variables is derived at its relevant planning level within a framework of detailed analysis and hence transmitted to the subsequent level, either upward or downward, to serve as either input or consistency check.

The decision variables within the hierarchical modeling system employed for the Pampanga Delta - Candaba Swamp Regional Development Project linking the three planning levels were defined as from bottom up as follows:

(i) Water resources development projects in terms of monthly distribution of water available for irrigated agriculture and the associated regime of water consumption of the irrigated lands within the Project Area.

(ii) Agricultural development plan consisting of land uses and cropping patterns consistent with the water available from an optimal combination of water resources projects and the simulated irrigation regime derived at the preceding level. The agricultural plan is hence translated into physical amounts of production and further on into values of production by crops and agricultural sectors.

(iii) Regional macro-economic development plan which provides for a regional general equilibrium consistent with independent demographic and socio-economic projections and with the gross outputs of the specific agricultural sectors derived in the preceding level.

If the projected growth rates of population and income per capita require capital and labor resources which exceed the demand for these resources due to the agricultural and water resources development plans, in order to sustain an overall balanced regional economy, these projections have to be modified and the development rate will be reduced. On the other hand, if the resulting agricultural development plan may account for an accelerated overall regional development

in terms of income per capita and restructuring of the regional economy in terms of employment and value added, again the projections may be modified and a varied scenario of regional development may thus result.

In the same manner, modification of the level of the agricultural development plan may be derived, within the second planning level of the hierarchical system, to be consistent with a preassigned development rate of the regional economy.

3.2 Functional Formulation of the Hierarchical System

I. Lower Level - Basin-wide Resources Analysis

I.1 Synthetic rainfall model - from insufficient daily rainfall records synthetic rainfall series were computed employing Gumbel's transformation and uniform distributed random numbers.

I.2 Hydro and agro meteorological models

$$\begin{bmatrix}\text{Total available}\\\text{instantaneous}\\\text{discharge}\end{bmatrix} = F\begin{bmatrix}\text{"natural"} & \text{groundwater} & \text{Diver-} & \text{change}\\\text{discharge;} & \text{contribution;} & \text{sions;} & \text{in}\\ & & & \text{storage}\end{bmatrix}$$

II. Intermediate - Sectoral Level - Agricultural Development

II.1 Land Allocation Model

Max - Annual net returns

Subject to: Balance equations for monthly land uses
Crop mix limitations; land availability constraints
Water availability constraints - monthly and annually consistent with the previous level.

II.2 Selection Model of Development Alternatives

Max - Annual net returns from agricultural development plans
Subject to: Development plans constraints; Water constraints consistent with lower level results; Marketing constraints; Employment constraints.

III. Highest Level Macro-Economic Analysis

Population and income per capita projections.
Derivation of total final demand.

$$\begin{bmatrix}\text{Gross output at}\\\text{target year } t\end{bmatrix} = \begin{bmatrix}(I-A)^{-1}\end{bmatrix}\begin{bmatrix}\text{Total final}\\\text{demand}\end{bmatrix}$$

Comparison of outputs for agricultural sectors derived here with production levels derived in intermediate level.

4. FORMULATION OF THE PLANNING MODELS OF THE HIERARCHICAL SYSTEM

4.1 The Project Development Level - The Hydro and Agro-Meteorological Simulation System

4.1.1 The Overall Simulation System

At the lower level of hierarchical system, the project development level, the determination of the options for water resources development has been carried out by a system of hydro- and agro-meteorological simulation models. The objectives of this phase were to:

(i) derive a water balance for the entire river basin and for its separate components (sub-basins) under theoretical, undisturbed conditions;

(ii) generate corresponding data on flood peaks and return periods;

(iii) establish diversion withdrawal and storage combinations for agricultural and other uses;

(iv) verifying various reservoir operations combinations within the basin system, so as to establish a consistent storage system for multi-period (30 years) means of natural conditions and man-made interferences.

The system of simulation models is presented in a schematic form in Fig. 2.

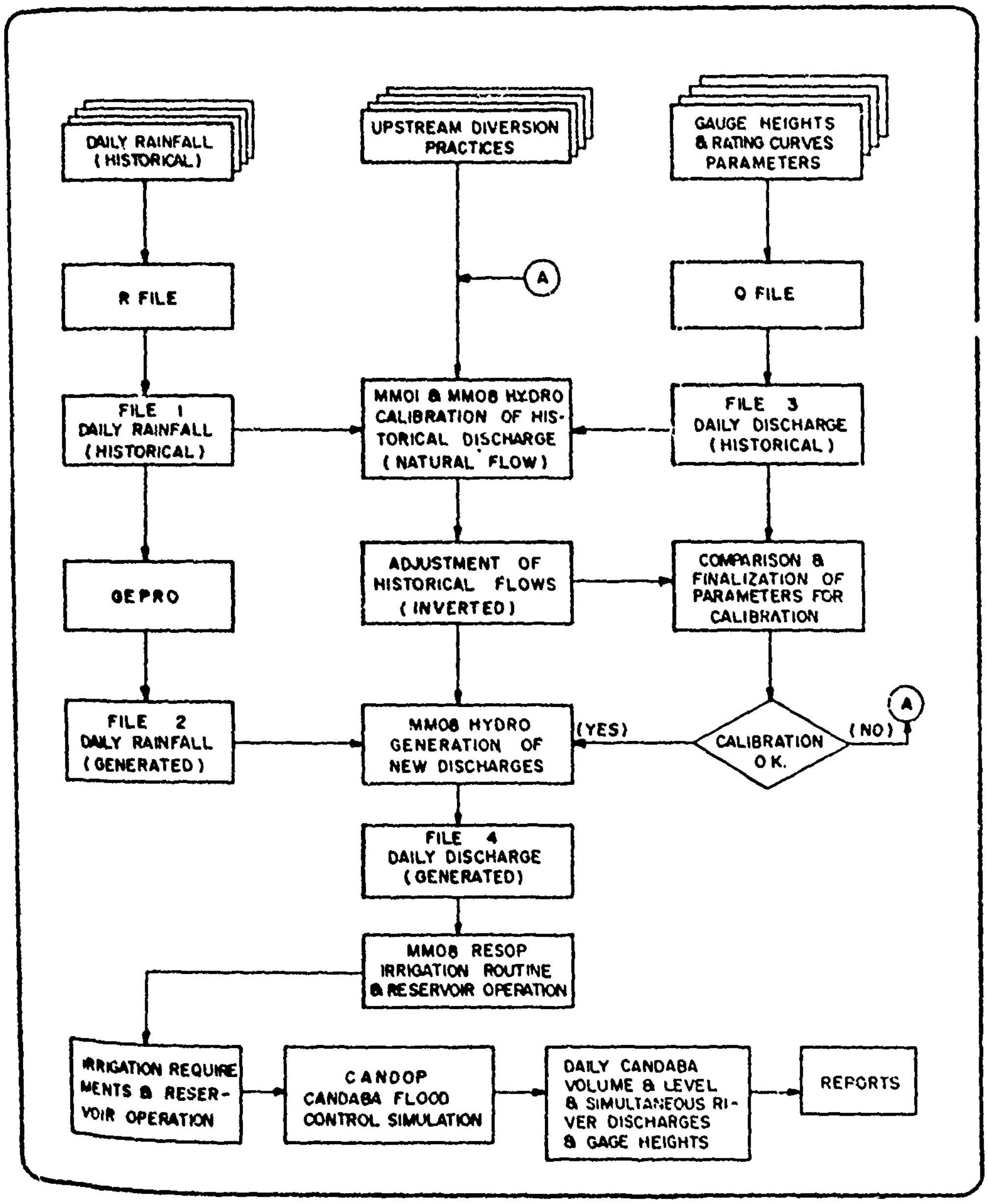
These models encompassed the entire river basin, simulating the flow regime in the river basin and in the Project Area on a probabilistic basis. The models employed river discharge data, basin characteristics, daily precipitation and other climatological and physical factors. They resulted in the derivation of alternative flow regimes under various man-made interferences and providing for the beneficial use of water for irrigation and for flood protection of agricultural and urban lands. The models determined - sequentially-water balances and identified the hydrological-cum-hydraulic feasibility of development projects for flood control and irrigation. They also determined water availability at various storage options for different levels of regional irrigated agricultural development.

4.1.2 The Hydro-meteorological Simulation Sub-system

The hydro-meteorological analysis is based on a systematic day-by-day moisture accounting procedure, developed by Mero [9] and referred to as the MM08-Hydro/Model; this analysis

FIG. 2

PAMPANGA DELTA/CANDABA SWAMP AREA
HYDRO-AGRO METEOROLOGICAL SYSTEM



follows the principles of the natural hydrologic cycle. The various model parameters, used as inputs for each of the sub-basins, were derived from observed river discharge data; these data were also used to confirm the validity of the simulated discharges.

The model simulations were executed over the whole watershed which was sub-divided into 28 interdependent, or independent, sub-basins. This sub-division was indicated by the availability of streamflow and other data.

The output of the studies served as primary reference data in the derivation of the theoretical (i.e. natural, unoperated, undiverted and/or unstored) water resource potential of the basin and the quantitative establishment of the dependence of the downstream areas (i.e. the project area) on the combined, simultaneous upstream operations:

- under undistrubed, natural conditions;
- under actual and eventual, future conditions, resulting from diversions, storage, etc.

4.1.2 The Agro-meteorological and Reservoir Operation Simulation Sub-systems

Following the generation of synthetic rainfall data and the generation of the "natural" undiverted river flows with the aid of the hydro-meteorological model, simulation of the entire river basin under existing conditions, i.e. diversion and withdrawal of water for agricultural and other uses, was performed.

This simulation, performed by the agro-meteorological model developed by Mero and Gilboa [10] and referred to as the MM08 RESOP Model, executes simultaneous calculations for daily water balances for up to ten varieties of crops per sub-basin or per water source, for present users upstream and for users downstream to be served by new projects.

The simulation model adheres closely to the full and detailed sequence of the hydrological cycle and the simultaneous agro-meteorological and reservoir operation water balances.

Both the agro-meteorological and reservoir operation cycles begin with the water input into their respective systems by means of natural precipitation and natural undiverted flows. Water losses are attributed to evaporation and surface and groundwater runoffs.

Given the area of the basin, the irrigable lands are defined together with a variety of possible crops. Each of these crops has a characteristic growth pattern with varying water requirements, dependent on such factors as evapotranspiration, soil-water-plant constants and soil type. The

irrigation requirements, on the other hand, were determined from water balance-type computations, in which plant and soil water requirements were balanced against rainfall occurrences.

Demand is satisfied either by rainfall, by river diversions, or by supply from storage reservoirs.

For the reservoir operation routine, the engineering characteristics of the existing and potential reservoirs were incorporated as input data into the reservoir operation routine. These data included: area-capacity curves, spillway crest elevations and lengths, bottom outlet elevations and areas, and reservoir operation rules according to the prime reservoir objectives, i.e., irrigation, flood control or power generation.

Testing varying reservoir combinations within the basin system and the analysis of their impact downstream in terms of flood peaks were major concerns of the study.

4.2 Optimization Models for Agricultural Land Use, Cropping Patterns and Project Selection

4.2.1 The Planning Process-Derivation of Alternative Agricultural Plans

At the second level, a two phase system of optimization models (linear programming) was employed.

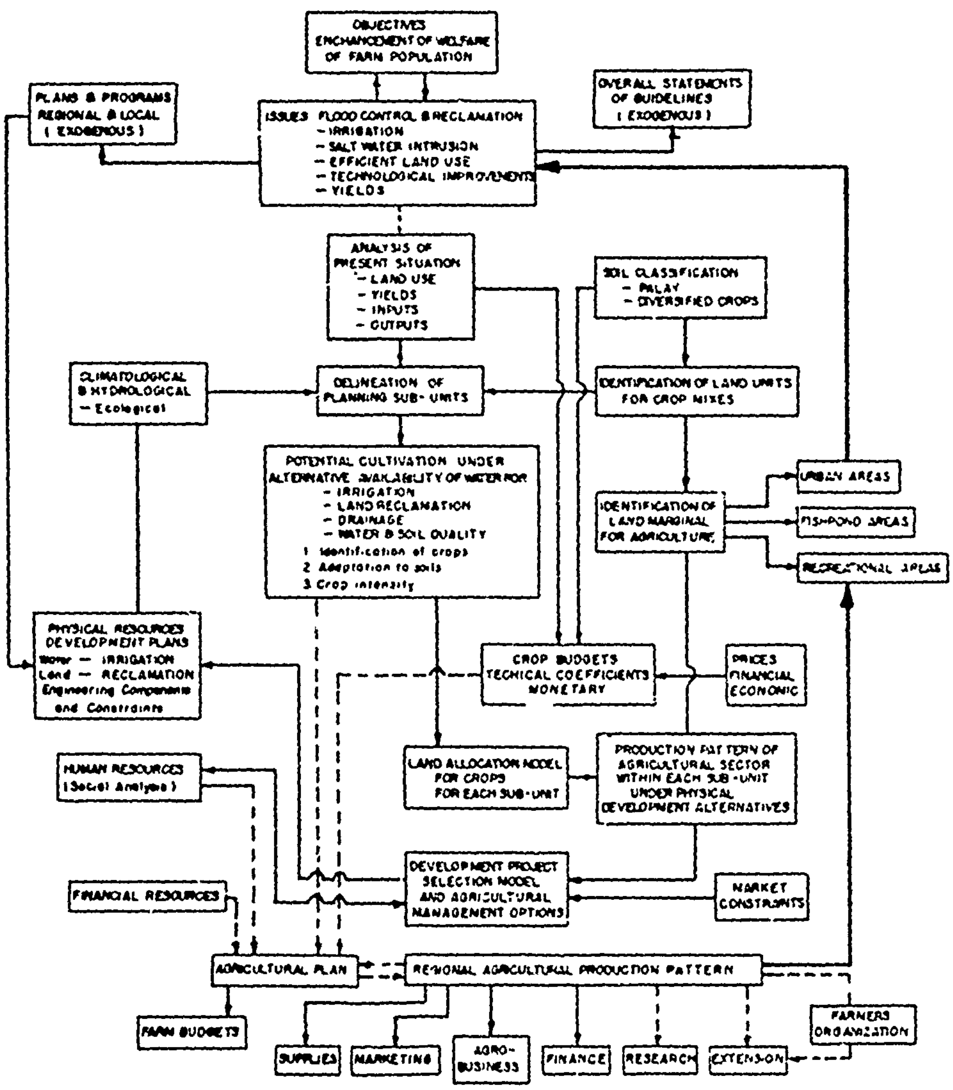
Due to the high degree of heterogeneity of the different sections of the project area, the simultaneous overall planning approach for agriculture was waived in favor of a two-level hierarchical approach which provides, at the first level, alternative agricultural plans for several planning sub-units which are affected by: input-output relations at the farm-gate level; a range of water availabilities for irrigation; crop intensity and crop rotation constraints; and the scope of potential improvements with regard to land reclamation and upgrading of soil productivities.

At the second level of the hierarchy, the selection of alternative development plans was done by taking into account regional economic factors (economic prices); project area - wide marketing and labor constraints; water availability constraints at potential sources; and possible interdependencies with regard to development and/or upgrading projects for certain sub-units.

The combinations of development plans for the various sub-units were evaluated at thisphase with regard to the physical development potential of the water resource sector of the area in terms of technological feasibility and economic justification. This planning phase yielded an economically-justified physical development plan with a corresponding agricultural production plan.

FIG. 3

SCHEME FOR AGRICULTURAL PLANNING WITHIN A HIERARCHICAL REGIONAL DEVELOPMENT APPROACH



A schematic illustration of this approach is given in Fig. 3.

4.2.2 The Land Use and Cropping Pattern Optimization Model

As stated above a land use model was formulated on the first level for each planning unit. The physical area of each pre-delineated planning unit was classified into soil suitability (for rice) and capability (for other crops) classes. For each soil class, two estimates of area were derived; without land reclamation projects (present) and with reclamation projects (future potential).

For each soil class, the potential crops were identified and the cropping calendar and input-output relations established. For this level, a linear programming land allocation model was separately formulated for each planning unit; this model derives optimal land uses under different assumptions of water availability and reclamation project costs and the associated trade-offs between the various options.

Constraints are formulated: for the projected cropping calendar, i.e., the months within a year during which the crops occupy land; for permissible portions of the specified soil classes to be occupied by a given crop due to crop rotation considerations; and for projected technological progress, diversification requirements and agro-technical constraints. In addition, maximum crop intensities are specified for irrigated and non-irrigated crops.

Parameters include net return coefficients which constitute the objective function, assumed cost figures for the improvement project, and monthly and annual water requirements for irrigation. Parametric changes in the assumed availability of water serve as the governing constraints for the extent of transformation from a rainfed cropping pattern to intensive, irrigated agriculture.

The objective function is formulated in terms of maximization of the annual net returns from agricultural production of a planning unit, i.e., net income from crop cultivation less annual costs due to the improvement projects.

Alternative development plans were derived for all planning units for different levels of investments, water availabilities and other assumptions.

The major, most valid, alternatives were screened out and used in the second level of the planning hierarchy.

4.2.3 Overall Agricultural Development and Project Selection Optimization Model

This second level of the planning hierarchy constitutes another linear programming model where the decision variables

are the alternative agricultural plans for each planning unit, encompassing the full range of potential water availability (zero to maximum) and extreme options of investment requirements for flood control and drainage (land improvement activities). The second set of decision variables consists of specific potential projects which would facilitate agricultural development of the area, such as single or multi-purpose reservoirs, groundwater development, and flood control schemes again at varying levels of effectiveness.

Constraints formulated for this model are of three types - resource availability constraints (water, employment), marketing constraints, and balance and dependency constraints.

The entire model is aimed at maximizing net returns on a regional basis, i.e., the net return from each plan in terms of economic (regional) prices, less annual outlays in investments, resulting from the projects to be selected in the optimization process.

The results of this model, inclusive of trade-off analysis, provided the agricultural development plan, i.e., cropping pattern and associated water resource projects, recommended for the project area, as well as the marginal analyses and trade-offs of deviations from this development option.

4.3 The Regional Macro-Economic Model

4.3.1 The Inter-industry System

The macro-economic regional analysis within the context of a multi-level approach to regional planning is the highest planning level; in the first iteration it forms the basis of the comprehensive regional plan. This highly aggregated level concerns itself with the major macro-economic determinants of the regional economy, namely, gross regional domestic product, personal income and expenditure, government expenditures, capital formation, trade balance and gross output. Together with the breakdown of these determinants according to various industrial sectors, the following analyses were derived:

- Structural analysis of the regional economy, inclusive of identification of bottlenecks, overall capital and employment ratios;
- Impact and multiplier analyses relating to income, capital, and employment, as well as such primary factors as land and water;
- Projecting overall and sectoral demands, i.e., levels of output, value added, and resource requirements.

At this highest level, a regional macro-economic accounting system was established employing the theory and practice of input-output models. Demographic and socio-economic projections were incorporated, and a projected quantitative scenario of the multi-sector regional economy was derived. The resulting projected gross regional domestic product and structural multipliers were tested for conformity and consistency with the outputs of the various agricultural subsectors which were obtained at the second level.

The implications of these results and analyses for the rest of the economy were then analyzed.

In practice a regional Inter-Industry (Input/Output) Model was constructed to permit macro-economic and sectoral analysis. The Input-Output Analysis concept was culled from the interest of economists in the forces that brought about a general equilibrium within the economy, with economic interdependence, and with the structure of the economy and the way in which the individual sectors fit together (Miernyk [11].

The Inter-Industry Accounts disaggregates the economy: on the one hand, into a processing segment which has a set of industrial sectors each producing one commodity or a group of homogeneous commodities, and on the other, into a final demand segment representing household and government expenditures, private investments and exports. This disaggregation permits the tracing of the flow of goods and services from one producing industry to another sector or industry that makes use of the commodity. In effect, the system shows how the output is distributed to the different industries and at the same time it shows the inputs each industry receives from other industries and sectors.

The input-output system is not a planning tool but rather an analytical tool that shows in detail how changes in one or more sectors of the economy will affect the total economy. The major advantage it brings to comprehensive planning is precisely this - the empirical demonstration of the interdependence of the different sectors within the economy. For normative planning, input-output analysis may be applied in conjunction with optimization techniques (Dorfman et al [12], Bargur [13]). However, this approach was not applied at this phase.

The Input-Output Accounts as an analytical tool brings with it several other forms of analysis (Chenery [14], such as structural analysis, forecasting and impact or multiplier analysis.

4.3.2 Empirical Application (Bargur [15])

Two computational algorithms to generate a regional input-output model from a national Input-Output table and to manipulate this model for analytical purposes by means of matrix calculations were derived for the project and employed in the sequence. The models are presented in Fig. 4.

5. ANALYSIS AND SUMMARY OF RESULTS

5.1 Formulation of the Regional Development Plan

Principal results of the models which correspond to the planning levels are shown in Fig. 5. They are presented in terms of the hierarchy of the planning concepts discussed in the preceding sections and summarize the components of the preferred development plan.

The results which pertain to the Lower Level Basin-Wide Water Resource Analysis, summarize at first the mean characteristics derived from the hydro-meteorological simulation. These in turn served as the basis for the derivation of four basin wide storage options with corresponding diversions, irrigated areas and flood peak reductions as well as four options of flood control for the Candaba Swamp, a major flood retention basin within the Project Area.

The Intermediate Sectoral Project Area Level sums up benefits and investments for the preferred plan which were derived from the agricultural development analyses for five planning units separately (crop intensities for each planning unit before and after the development represent the direct gains or benefits) and from the various storage and flood control options established at the first, lower level of the hierarchical system.

The preferred plan has been characterized by additional irrigated area, flood relieved area, additional water supply and the corresponding investment agricultural production and benefits (income and job opportunities).

Finally the highest level of the planning system which pertains to the macro-economic analysis provides the overall regional reference in which the agricultural development will take place.

Regional growth rates for 30 years for the gross regional product, value added, employment production and capital investments have been derived so as to sustain corresponding growth rates for the agricultural sector as derived in the intermediate planning level.

FIG. 4

SYSTEM OF MODELS AND PROCEDURES FOR REGIONAL INPUT-OUTPUT ANALYSIS

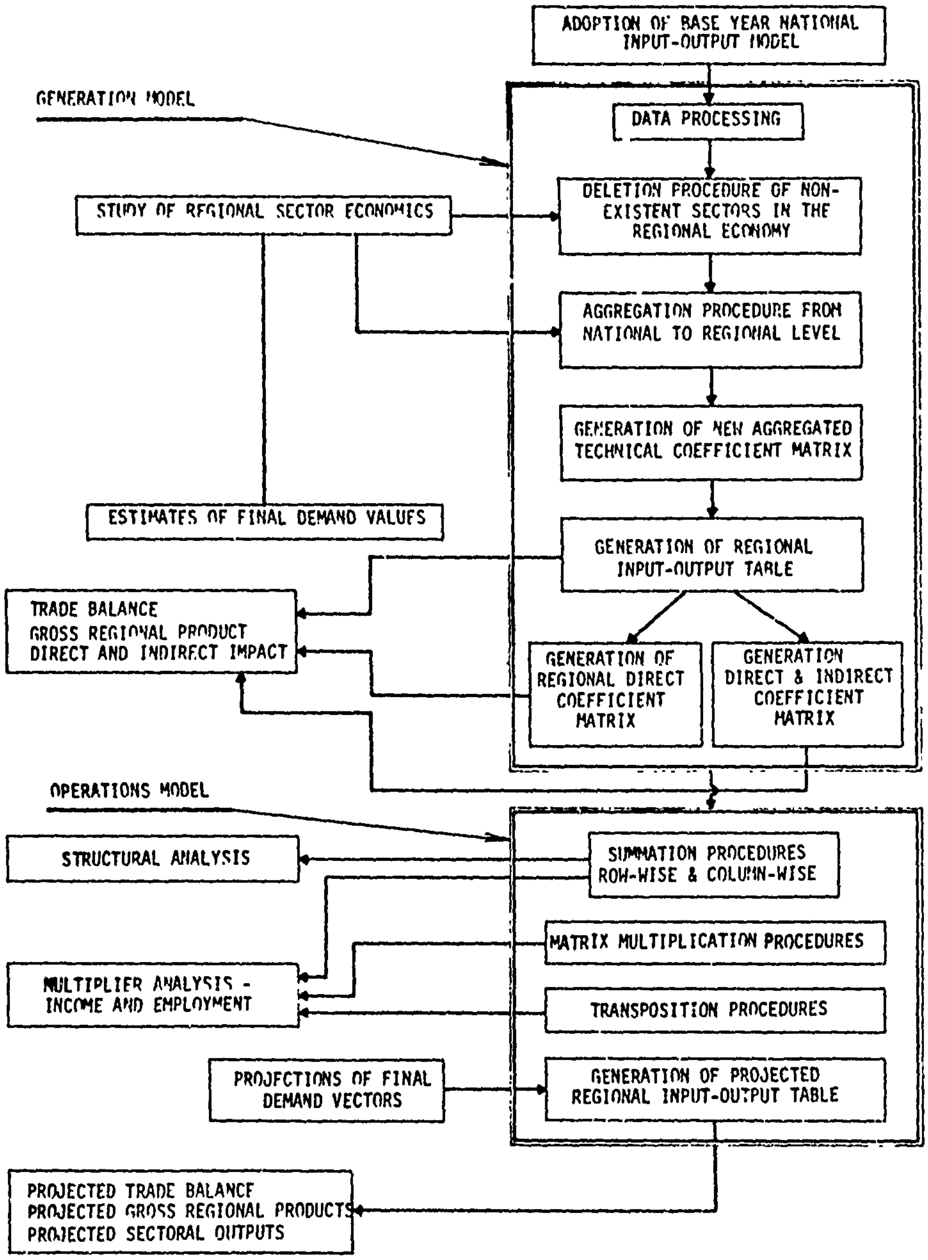
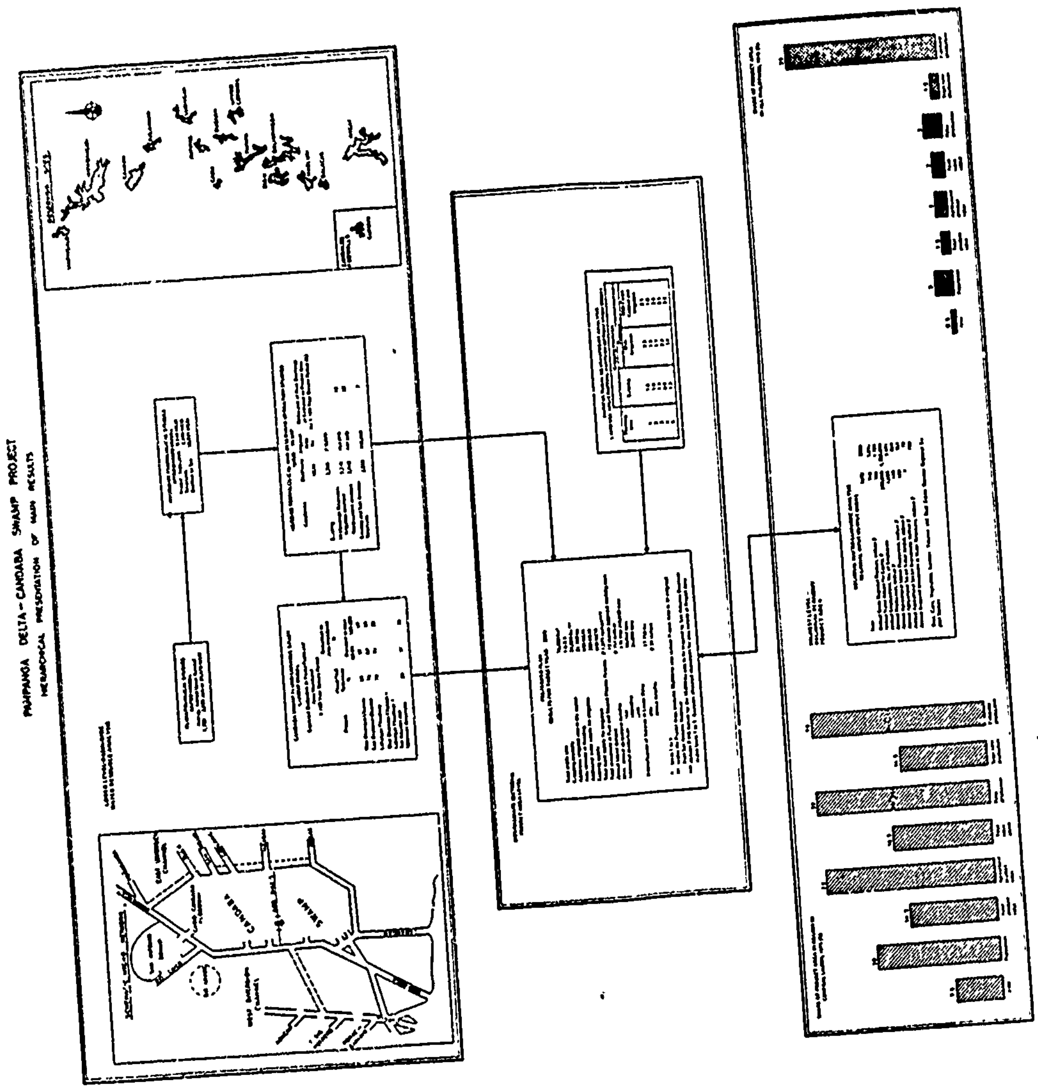


FIG. 5


PAMPANGA DELTA-CANDABA SWAMP PROJECT
HEIRARCHICAL PRESENTATION OF MAIN RESULTS

5.2 Assessment

The economic and socio-economic consistency between o overall regional and sectoral (agricultural) development has been checked by inter-level comparison of the intermediate and highest planning levels.

This consistency check corresponds to the technological consistency check applied by inter-level feedback between the lowest and intermediate planning levels.

Hence the hierarchical planning system as presented here and practically employed could actually be viewed as a decomposition system where technological feasibility, economic efficiency and socio-economic benefits are analyzed independently but tied together by the inter-level decision-making components.

This independent, yet simultaneous approach to the techno-economic and socio facets of a regional development plan, has proven itself as an applicable and useful tool in regional development practice.

5.3 Problems and Limitations

One major problem encountered in this project stems mainly from incompatibility of the area delineated for study. Projects are usually delineated on the demand side along administrative and geo-political lines, such as administrative regions, states provinces, metropolitan areas and municipalities. However, on the supply side, i.e. the resource base, only geo-physical entities such as river basins - are valid. Regional plans, as dealt with here, may thus create jurisdictional problems, while at the same time the statistical data necessary for regional delineation is lacking. These problems are usually compounded by incompatible inter-regional and national development goals such as metropolitan extension or inter-regional competition of the allocation of capital funds, as also by the usually competitive interrelations of the political and administrative authorities.

All these problems were actually magnified in the Pampanga Delta-Candaba Swamp Project. However, application of the hierarchical approach provided, in addition to the above-mentioned theoretical arguments, a decentralized planning system which could assume a varied pace of progress for the different planning levels according to the availability of data without losing the overall comprehensiveness of the regional plan, as long as the inter-level linkages were preserved throughout the process.

5.2 Assessment

The economic and socio-economic consistency between o overall regional and sectoral (agricultural) development has been checked by inter-level comparison of the intermediate and highest planning levels.

This consistency check corresponds to the technological consistency check applied by inter-level feedback between the lowest and intermediate planning levels.

Hence the hierarchical planning system as presented here and practically employed could actually be viewed as a decomposition system where technological feasibility, economic efficiency and socio-economic benefits are analyzed independently but tied together by the inter-level decision-making components.

This independent, yet simultaneous approach to the techno-economic and socio facets of a regional development plan, has proven itself as an applicable and useful tool in regional development practice.

5.3 Problems and Limitations

One major problem encountered in this project stems mainly from incompatibility of the area delineated for study. Projects are usually delineated on the demand side along administrative and geo-political lines, such as administrative regions, states provinces, metropolitan areas and municipalities. However, on the supply side, i.e. the resource base, only geo-physical entities such as river basins - are valid. Regional plans, as dealt with here, may thus create jurisdictional problems, while at the same time the statistical data necessary for regional delineation is lacking. These problems are usually compounded by incompatible inter-regional and national development goals such as metropolitan extension or inter-regional competition of the allocation of capital funds, as also by the usually competitive interrelations of the political and administrative authorities.

All these problems were actually magnified in the Pampanga Delta-Candaba Swamp Project. However, application of the hierarchical approach provided, in addition to the above-mentioned theoretical arguments, a decentralized planning system which could assume a varied pace of progress for the different planning levels according to the availability of data without losing the overall comprehensiveness of the regional plan, as long as the inter-level linkages were preserved throughout the process.

LIST OF REFERENCES

[1] Klahr, C.N., MULTIPLE OBJECTIVES IN MATHEMATICAL PROGRAMMING, Operations Research, 6 (1958).

[2] Nijkamp, P., THE USE OF SPATIAL IMPACT MODELS IN PHYSICAL PLANNING , Geoforum 7 (1976).

[3] Nijkamp, P., A MULTI-CRITERIA ANALYSIS FOR PROJECT EVALUATION; ECONOMIC-ECOLOGICAL EVALUATION OF A LAND RECLAMATION PROJECT, Regional Science Association 35 (1975),

[4] Delf A. Van and P. Nijkamp, THE USE OF HIERARCHICAL OPTIMIZATION CRITERIA IN REGIONAL PLANNING, Journal of Regional Science, Vol. 17, No. 2, Aug. 1977.

[5] Haimes, Y.Y., HIERARCHICAL MODELING OF REGIONAL TOTAL WATER RESOURCES SYSTEMS, Automatica, Vol. II, Pergamon Press, (1975).

[6] Mesarovic, M.D., D. Macko and T. Takahara, THEORY OF HIERARCHICAL MULTI-LEVEL SYSTEMS, Academic Press, N.Y. (1970).

[7] Haimes, Y.Y., W.A. Hall and H.T. Freedman, MULTI-OBJECTIVE OPTIMIZATION IN WATER RESOURCES SYSTEMS: THE SURROGATE WORTH TRADE-OFF METHOD, Elsevier, The Netherlands, (1975).

[8] Bargur, J., THE PAMPANGA DELTA-CANDABA SWAMP DEVELOPMENT PROJECT - AN EFFORT IN COMPREHENSIVE REGIONAL PLANNING, Philippine Planning Journal, Vol. II, No. 2 (April 1976).

[9] Mero, F., AN APPROACH TO DAILY HYDRO-METEOROLOGICAL WATER BALANCE COMPUTATIONS FOR SURFACE AND GROUNDWATER BASINS, Proceedings ITC-UNESCO Seminar for Integrated River Basin Development, (1969).

[10] Gilboa, Y. and F. Mero, THE BUTUCATU AQUIFER OF SOUTH AMERICA, MODEL OF AN UNTAPPED CONTINENTAL AQUIFER, Journal of Hydrology (1976).

[11] Miernyk, W., ELEMENTS OF INPUT-OUTPUT ANALYSIS, Random House, N.Y. (1976).

[12] Dorfmann, R., P.A. Samuelson and R.M. Solow, LINEAR PROGRAMMING AND ECONOMIC ANALYSIS, McGraw Hill, N.Y. (1958).

[13] Bargur, J., A DYNAMIC MULTI-SECTOR PROGRAMMING APPROACH TO REGIONAL WATER RESOURCES MANAGEMENT, Water Resources Research, Vol. 8, No. 4 (1971).

[14] Chenery, Hollis B. and Paul C. Clark, INTER-INDUSTRY ECONOMICS, John Wiley and Sons Inc., N.Y. (1959).

[15] Bargur, J., COMPUTATIONAL ASPECTS OF INPUT-OUTPUT ANALYSIS APPLIED TO RESOURCES MANAGEMENT, SERL Report No. 70-7, University of California, Berkeley, California (1970).

논 문 집

-태평양지역운영분석회의-

(1979, 4, 23 - 28 : 서울)

제 II 권

THE PACIFIC CONFERENCE
1979

주 최

한국군사운영분석연구회

한 국 운 용 과 학 회

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.